TSAR Project Part 1

Michèle Fille

March 28, 2024

Table of contents

List of Figures

List of Tables

Preliminaries

```
# Install necessary packages for this quarto file
#install.packages("data.table") # already installed
#install.packages("dplyr") # already installed
#install.packages("knitr") # already installed
#install.packages("ggplot2") # already installed
#install.packages("GGally") # already installed
#install.packages("plotly") # already installed
#install.packages("skimr") # already installed

# Load necessary libraries for this quarto file
library(data.table)
library(dplyr)
library(skimr)
library(knitr)
library(ggplot2)
library(GGally)


# Read the data
LCdata <- fread("LCdata.csv")

# Sample the data
set.seed(4)
myLCdata <- LCdata %>% sample(5) %>% slice_sample(prop = .5)

# Save myLCdata as CSV file (for project part 2)
fwrite(myLCdata, "myLCdata.csv")
```

Before starting the analysis, it was ensured that the necessary tools and data were available. The essential R packages were installed and loaded. Please make sure to run this R Code Chunk first, before trying to run the following ones.

The Lending Club (`LCdata`) data was then read from the provided CSV file named "LCdata.csv" using the `fread()` function.

In preparation for the exploratory analysis, a subset of the data set (`myLCdata`) was selected as a sample. The TSAR ID was used as the seed value (in this case seed value 4) to ensure the reproducibility of the results. The code randomly selected 5 out of 19 attributes and 50% of the rows.

The above R Code Chunk illustrates these first steps.

Task 1. Business Understanding

```
# Display a summary of your dataset using glimpse() and str()
glimpse(myLCdata)
```

```
Rows: 44,369
Columns: 5
$ mths_since_last_delinq <int> 13, 39, 7, NA, 42, NA, 6, NA, 20, NA, NA, 33, N~
$ emp_length             <chr> "< 1 year", "1 year", "2 years", "10+ years", "~
$ purpose                <chr> "debt_consolidation", "debt_consolidation", "ca~
$ open_acc               <int> 14, 16, 13, 12, 11, 12, 9, 11, 18, 10, 8, 19, 8~
$ tot_cur_bal            <int> 19086, 22487, 11068, 204179, 260166, 54524, 183~
```

```
str(myLCdata)
```

```
Classes 'data.table' and 'data.frame':  44369 obs. of  5 variables:
 $ mths_since_last_delinq: int  13 39 7 NA 42 NA 6 NA 20 NA ...
 $ emp_length            : chr  "< 1 year" "1 year" "2 years" "10+ years" ...
 $ purpose               : chr  "debt_consolidation" "debt_consolidation" "car" "credit_card" ...
 $ open_acc              : int  14 16 13 12 11 12 9 11 18 10 ...
 $ tot_cur_bal           : int  19086 22487 11068 204179 260166 54524 183414 8964 531074 218908 ...
```

The `glimpse()` and `str()` functions provide a first impression of the data sample (`myLCdata`). The `glimpse()` function indicates that there are 44,369 rows (data objects = individual loan records) and 5 columns (attributes). The `str()` function displays the structure of the data frame, including the data types of each column and a preview of the first values.

An overview of the attributes can be found in the following Table 1, the descriptions are taken from the file LD_DataDictionary.xlsx (downloaded from Moodle), the attribute names and types from the R Code Chunk above.

Table 1: Attributes from data set myLCdata, their type and the Description.

Attribute Name	Attribute Type	Description of Attribute
emp_length	chr	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
mths_since_last_delinq	int	The number of months since the borrower's last delinquency.
open_acc	int	The number of open credit lines in the borrower's credit file. Remark: - A credit line is a type of loan that allows an individual or business to borrow money and repay it, often on a revolving basis without applying for a new loan. - A revolving credit is a credit line that remains available even as you pay the balance. Borrowers can access credit up to a certain amount and then have ongoing access to that amount of credit. Typically a credit card.
purpose	chr	A category provided by the borrower for the loan request.
tot_cur_bal	int	Total current balance of all accounts.

Task 2. Data Understanding Based on Summary Statistics

```
# Review main statistical metrics of attributes using summary()
summary(myLCdata)
```

```
mths_since_last_delinq  emp_length          purpose             open_acc
Min.   :  0.00         Length:44369       Length:44369       Min.   : 0.00
1st Qu.: 15.00         Class :character   Class :character   1st Qu.: 8.00
Median : 31.00         Mode  :character   Mode  :character   Median :11.00
Mean   : 33.99                                               Mean   :11.54
3rd Qu.: 50.00                                               3rd Qu.:14.00
Max.   :143.00                                               Max.   :76.00
NA's   :22776                                                NA's   :3
 tot_cur_bal
Min.   :      0
1st Qu.:  29561
Median :  79916
Mean   : 138467
3rd Qu.: 207158
Max.   :4127799
NA's   :3427
```

```
# Convert "n/a" to NA in emp_length column
myLCdata$emp_length <- na_if(myLCdata$emp_length, "n/a")

# Review main statistical metrics of attributes using skim()
skim(myLCdata)
```

Table 2: Summary statistics for characters and numeric variables

(a) Data Summary

Name	myLCdata
Number of rows	44369
Number of columns	5
Key	NULL
Column type frequency:	
character	2
numeric	3
Group variables	None

Variable type: character

(b) Summary of character variables

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
emp_length	2232	0.95	6	9	0	11	0
purpose	0	1.00	3	18	0	14	0

Variable type: numeric

(c) Summary of numeric variables

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100
mths_since_last_delinq	22776	0.49	33.99	21.88	0	15	31	50.0	143
open_acc	3	1.00	11.54	5.37	0	8	11	14.0	76
tot_cur_bal	3427	0.92	138467.17	151513.69	0	29561	79916	207157.8	4127799

Before using the `skim()`function, it must be made sure that all missing values are recognized. One unique value of emp_length attribute "n/a", what gets not automatically recognized as NA. Before the "convert line" was inserted, the `skim()`function did show 0 n_missing values, but as Table 3 shows there are indeed some. Therefore I added this line of code, to make sure the missing values for emp_length are recognized as such.

The `summary()` function cannot provide further insights for variables with the type of character, whereas the `skim()` function can, see Table 2, especially Table 2b . For numeric variables both functions show the same, however the `skim()` function gives some additional information, such as the `complete_rate`, the standard deviation (`sd`), and a histogram (`hist`), which can be seen in Table 2c. Further more the `skim()` function also provides a nice overview of the whole data, see Table 2a.

In the following each of the five attributes (Table 1) will be interpreted individually.

Employment Length

This attribute denotes the employment length of the borrower in years.

- As it is a character variable, summary statistics such as mean and quartiles are not applicable.
- There are 2,232 missing values and the attribute has a `complete_rate` of 95%, as shown in the Table 2 produced by the `skim()`function. This says that about 5% of the values are missing, what is confirmed in Table 3 and indicates completeness of the data.
- The min (6), max (9), and n_unique (11) indicate the range of unique values present in the emp_length attribute. The min and max values represent the shortest and longest lengths among these unique values, respectively. In this case, the minimum length is 6 characters, and the maximum length is 9 characters. The value of n_unique indicates that there are 11 unique values - excluding NAs - present for the attribute emp_length.

  The following R Code Chunk shows the unique values of emp_length, unsing the `unique()` function. This reveals the 11 unique values plus the value NA. The code output indicates that category with the min characters of 6 must be "1 year" and the one with the max characters of 9 must be "10+ years".

```
# Display unique values
sort(unique(myLCdata$emp_length), na.last = TRUE)
```

```
[1] "< 1 year"  "1 year"    "10+ years" "2 years"   "3 years"   "4 years"
[7] "5 years"   "6 years"   "7 years"   "8 years"   "9 years"   NA
```

- To check if the value frequencies are unbalanced, it needs further investigation see the R Code Chunk below and its output in Table 3.

```r
# Calculate counts for each category of emp_length, including NAs
emp_length_counts <- table(myLCdata$emp_length, useNA = "always")

# Sort the categories
sorted_emp_length <- c("< 1 year", "1 year", "2 years", "3 years", "4 years", "5 years", "6 years",

# Reorder the counts based on the sorted categories
emp_length_counts <- emp_length_counts[match(sorted_emp_length, names(emp_length_counts))]

# Calculate percentages for each category
emp_length_percentages <- paste0(round(prop.table(emp_length_counts) * 100, 2), '%')

# Combine counts and percentages into a data frame
emp_length_summary <- data.frame(
  emp_length = names(emp_length_counts),
  count = as.vector(emp_length_counts),
  percentage = emp_length_percentages)

# Print the table using kable
knitr::kable(emp_length_summary)
```

Table 3: Frequencies of emp_length values - absolute counts and as percentages

emp_length	count	percentage
< 1 year	3586	8.08%
1 year	2744	6.18%
2 years	3897	8.78%
3 years	3496	7.88%
4 years	2597	5.85%
5 years	2769	6.24%
6 years	2210	4.98%
7 years	2289	5.16%
8 years	2223	5.01%
9 years	1767	3.98%
10+ years	14559	32.81%
NA	2232	5.03%

- Table 3 shows the discrepancy in counts among the categories indicates an imbalance in the distribution of employment lengths. Specifically, there is a higher prevalence of borrowers with longer employment lengths (10+ years: 14,559 occurrences, 32.81% of all occurrences) compared to those with shorter employment lengths (e.g., < 1 year: 3,586 occurrences, 8.08% of all occurrences, or 1 year: 2,744 occurrences, 6.18% of all occurrences).
- In the domain context, a longer period of employment could indicate career advancement and a higher income. Additionally, individuals with longer employment periods are likely to be older and have accumulated more experience in the job market. This may signal greater financial stability and reliability, potentially reducing the perceived credit risk for lenders. Similarly, longer periods of employment could indicate a stronger credit profile, which could increase the likelihood of lending to these individuals.

Months Since Last Delinquency

This attribute represents the number of months since the borrower's last delinquency.

- The minimum value is 0, suggesting that some borrowers have had delinquencies fairly recently.
- The maximum value is 143, indicating that some borrowers have not had any delinquency.
- The median (31) and mean (33.99) indicate that, on average, there is a considerable period since the last delinquency.

```
# Calculate the Interquartile Range (IQR) considering missing values
q1 <- quantile(myLCdata$mths_since_last_delinq, 0.25, na.rm = TRUE)
q3 <- quantile(myLCdata$mths_since_last_delinq, 0.75, na.rm = TRUE)
iqr <- q3 - q1

# Define a threshold for identifying outliers (1.5 times the IQR)
threshold <- 1.5

# Identify outliers
lower_bound <- q1 - threshold * iqr
upper_bound <- q3 + threshold * iqr
outliers <- myLCdata$mths_since_last_delinq[myLCdata$mths_since_last_delinq < lower_bound | myLCdat

# Sort outliers in descending order
outliers <- sort(outliers, decreasing = TRUE)

# Create a formatted string of outliers
outliers_mths_since_last_delinq <- paste("{", paste(outliers, collapse = ", "), "}", sep = "")

# Output outliers
paste0("outliers of mths_since_last_delinq = ", outliers_mths_since_last_delinq)
```

```
[1] "outliers of mths_since_last_delinq = {143, 135, 133, 120, 116, 114, 108, 105, 105, 104}"
```

Interpretation of Outlier Threshold: In identifying outliers for the 'mths_since_last_delinq' attribute, a threshold of 1.5 times the interquartile range (IQR) was employed. This threshold offers a moderate approach, capturing data points that moderately deviate from the central tendency of the distribution.

- The median (31) and the mean (33.99) are relatively close together; however, the maximum (143) is relatively high compared to the median and mean. This suggests that the distribution of months since the last delinquency is right-skewed. The histogram generated by the `skim()` function also indicates a right-skewed distribution for the attribute representing the number of months since the last delinquency. This skewness suggests that the majority of borrowers have relatively shorter periods since their last delinquency, with a smaller proportion experiencing longer periods. The proximity of the median and mean further supports this observation. However, the presence of outliers (see the R Code Chunk above), particularly the maximum value of 143 months, suggests that some borrowers have experienced significantly longer periods since their last delinquency. This skewness and the presence of outliers should be considered when interpreting this attribute, as it may have implications for the lending process. Further exploration, such as identifying and possibly removing outliers, could provide more insights into the distribution and potential implications for the lending process.
- The standard deviation (SD) of 21.88, from the `skim()`function, indicates the average deviation of individual data points from the mean of 33.99. A standard deviation of around 22 suggests that the distribution of the number of months since the last delinquency exhibits a moderate level of variability around the mean. This means that while most borrowers have delinquency periods clustered around the mean, there is also some variability, with a significant portion of them having periods either significantly shorter or longer than the mean. This variability underscores the diverse delinquency periods of borrowers, emphasizing the importance of understanding and considering such differences in the loan approval process.

- There are 22,776 missing values, which is significant and may indicate that these borrowers have no delinquency history or that the data is not available. Also the `complete_rate` of only 49% shows this moderate incompleteness in the data. Dealing with missing values is important for the data preparation step. In handling these missing values, it may be beneficial to add a category that indicates when a borrower never had a delinquency.
- In the domain context it could mean that individuals with short periods since their last delinquency may have turned to alternative sources of credit, such as this peer-to-peer lending platform, as this may have less stringent credit requirements than traditional lenders, such as banks.

Open Credit Lines

This attribute represents the number of open credit lines in the borrower's credit file.

- The minimum value is 0, suggesting that some borrowers do not have any open credit lines in their credit file.
- The maximum value is 76, indicating that some borrowers have fairly a lot of open credit lines in their credit file.
- The median (11) and mean (11.54) indicate that, on average, borrowers have some open credit lines in their credit files.

```r
# Calculate the Interquartile Range (IQR) considering missing values
q1 <- quantile(myLCdata$open_acc, 0.25, na.rm = TRUE)
q3 <- quantile(myLCdata$open_acc, 0.75, na.rm = TRUE)
iqr <- q3 - q1

# Define a threshold for identifying outliers (6 times the IQR)
threshold <- 6

# Identify outliers
lower_bound <- q1 - threshold * iqr
upper_bound <- q3 + threshold * iqr
outliers <- myLCdata$open_acc[myLCdata$open_acc < lower_bound | myLCdata$open_acc > upper_bound]

# Sort outliers in descending order
outliers <- sort(outliers, decreasing = TRUE)

# Create a formatted string of outliers
outliers_open_acc <- paste("{", paste(outliers, collapse = ", "), "}", sep = "")

# Output outliers
paste0("outliers of open_acc = ", outliers_open_acc)
```

```
[1] "outliers of open_acc = {76, 70, 56, 56, 55, 55}"
```

Interpretation of Outlier Threshold: In identifying outliers for the 'open_acc' attribute, a threshold of 6 times the interquartile range (IQR) was utilized. This threshold represents an aggressive approach, capturing data points that deviate significantly from the central tendency of the distribution. By setting such a high threshold, it was aimed to identify extreme values that may have substantial implications for the lending process, such as borrowers with an unusually high number of open credit lines.

- The median (11) and the mean (11.54) are relatively close together; however, the maximum (76) is relatively high compared to the median and mean. This suggests that the distribution of open credit lines in the borrowers credit file is right-skewed. The histogram generated by the 'skim()' function also indicates a right-skewed distribution for this attribute. This skewness suggests that the majority of borrowers have relatively moderate amount of open credit lines, with a smaller proportion that has many more open credit lines. The proximity of the median and mean further supports this observation. However, the presence of outliers (see the R Code Chunk above), particularly the maximum value of 76 open credit lines, suggests that some borrowers have significantly more open credit lines. This skewness and the presence of outliers should be considered when interpreting this attribute, as it may have implications for the lending process. Further exploration, such as identifying and possibly removing outliers, could provide more insights into the distribution and potential implications for the lending.
- The standard deviation (SD) of 5.37, from the `skim()`function, indicates the average deviation of individual data points from the mean of 11.54. A standard deviation of around 5.37 suggests that the distribution of the number of open credit lines in the borrower's credit file exhibits a moderate level of variability around the mean. This means that while most borrowers have a moderate number of open credit lines clustered around the mean, there is also some variability, with a significant portion of them having a slightly lower or higher number of open credit lines. This variability underscores the diverse number of open credit lines among borrowers, emphasizing the importance of understanding and considering such differences in the loan approval process.
- There are only 3 missing values, which is insignificant, as also the complete_rate is (rounded) 100%, indicating completeness in the data.
- In the domain context, this could suggest that lenders perceive borrowers with a low to moderate number of open credit lines as more stable and financially responsible. Consequently, such borrowers would be more appealing candidates for obtaining a loan.

Purpose of Loan

This attribute describes the purpose for which the loan was requested.

- Similar to employment length, it is a character variable, so summary statistics such as mean and quartiles are not applicable.
- There are no missing values in the "purpose" attribute, indicating that each loan application has a stated purpose. However, the presence of a category "other" indicates that some applications may not fit neatly into the predefined categories - if they are predefined. While "other" cannot be technically considered as missing data, it still represents an unusual category. "Other" is employed to encompass purposes not explicitly listed or to bypass further specification of the specific purpose. The category "other" can include a wide range of reasons. Therefore, it may be helpful or even necessary to treat "other" as a form of missing value. Table 4 indicates that the category "other" was selected 2,050 times, constituting approximately 4.6% of the total. While this percentage is not excessively high, it still represents a significant portion of the data set.
- The min (3), max (18), and n_unique (14) indicate the range of unique values present in the purpose attribute. The min and max values represent the shortest and longest lengths among these unique values, respectively. In this case, the minimum length is 3 characters, and the maximum length is 18 characters. The value of n_unique indicates that there are 14 unique values (no NAs) for this attribute.

The following R Code Chunk shows the unique values of purpose, using the `unique()` function. This reveals the 14 unique values that the category with the min characters of 3 is "car" and the one with the max characters of 18 is "debt_consolidation".

```
# Display unique values
unique((myLCdata$purpose))
```

```
 [1] "debt_consolidation" "car"                "credit_card"
 [4] "home_improvement"   "other"              "small_business"
 [7] "medical"            "major_purchase"     "vacation"
[10] "house"              "moving"             "educational"
[13] "wedding"            "renewable_energy"
```

- To check if the value frequencies are unbalanced, it needs further investigation see the R Code Chunk below and its output in Table 4.

```
# Calculate counts for each category of purpose
purpose_counts <- table(myLCdata$purpose)

# Sort the categories based on counts in increasing order
sorted_purpose <- names(sort(purpose_counts))

# Reorder the counts based on the sorted categories
purpose_counts <- purpose_counts[match(sorted_purpose, names(purpose_counts))]

# Calculate percentages for each category
purpose_percentages <- paste0(round(prop.table(purpose_counts) * 100, 2), '%')

# Combine counts and percentages into a data frame
purpose_summary <- data.frame(
  purpose = names(purpose_counts),
  count = as.vector(purpose_counts),
  percentage = purpose_percentages)

# Print the table using kable
knitr::kable(purpose_summary)
```

Table 4: Frequencies of purpose values - absolute counts and as percentages

purpose	count	percentage
educational	23	0.05%
renewable_energy	36	0.08%
wedding	121	0.27%
house	176	0.4%
vacation	255	0.57%
moving	279	0.63%
medical	431	0.97%
car	467	1.05%
small_business	487	1.1%
major_purchase	852	1.92%
other	2050	4.62%
home_improvement	2632	5.93%
credit_card	10310	23.24%
debt_consolidation	26250	59.16%

- The value frequencies for the "purpose" attribute, see Table 4, shows a clear imbalance. Notably, "debt_consolidation" and "credit_card" constitute the majority of cases, representing 59.16% and 23.24% respectively. These categories typically indicate financial obligations or debt repayment, which may be considered less favorable. Conversely, categories like "educational" and "renewable_energy" are less frequent, accounting for only 0.05% and 0.08% respectively. These purposes suggest investments in education or environmentally friendly practices, which could potentially lead to greater financial stability or savings. The stark contrast between these categories underscores an imbalance in the data set, highlighting the prevalence of loan applications geared towards debt management over those aimed at self-improvement or sustainable practices. Such imbalances may introduce bias in analyses or predictions, necessitating careful consideration in any subsequent analysis or decision-making process.
- In the domain context, the prevalence of loan applications for purposes such as "debt_consolidation" and "credit_card" suggests that individuals utilizing this platform may be facing financial challenges or existing debt obligations. This pattern could imply that traditional financial institutions, like banks, may not perceive these individuals as creditworthy or trustworthy enough to extend further credit, leading them to seek alternative borrowing options. Consequently, they may resort to platforms like this one to consolidate their debts or manage existing financial burdens.

Total Current Balance

This attribute indicates the total current balance of all accounts.

- The minimum value is 0, suggesting that some borrowers do not have any money on all of their accounts, a balance of 0.
- The maximum value is 4,127,799, indicating that some borrowers have a high total balance of all their accounts.

```
# Calculate the Interquartile Range (IQR) considering missing values
q1 <- quantile(myLCdata$tot_cur_bal, 0.25, na.rm = TRUE)
q3 <- quantile(myLCdata$tot_cur_bal, 0.75, na.rm = TRUE)
iqr <- q3 - q1

# Define a threshold for identifying outliers (8 times the IQR)
threshold <- 8

# Identify outliers
lower_bound <- q1 - threshold * iqr
upper_bound <- q3 + threshold * iqr
outliers <- myLCdata$tot_cur_bal[myLCdata$tot_cur_bal < lower_bound | myLCdata$tot_cur_bal > upper_

# Sort outliers in descending order
outliers <- sort(outliers, decreasing = TRUE)

# Create a formatted string of outliers
outliers_tot_cur_bal <- paste("{", paste(outliers, collapse = ", "), "}", sep = "")

# Output outliers
paste0("Outliers of tot_cur_bal = ", outliers_tot_cur_bal)
```

```
[1] "Outliers of tot_cur_bal = {4127799, 3078704, 2629423, 2276535, 1907327, 1813088, 1728172, 166734(
```

Interpretation of Outlier Threshold: In identifying outliers for the 'tot_cur_bal' attribute, a threshold of 8 times the interquartile range (IQR) was utilized. This threshold represents an aggressive approach, capturing data points that deviate significantly from the central tendency of the distribution. By setting such a high threshold, it was aimed to identify extreme values that may have substantial implications for the lending process, such as borrowers with an unusually high total balance on all their accounts.

- The median (79,916) and mean (138,467) are relatively far apart. The mean value being significantly lower than the mean suggests a potential right-skewed distribution. This indicates that while most borrowers have relatively lower total balances, a small proportion may have substantially higher balances, skewing the mean upwards. This observation is supported by the presence of outliers, particularly the maximum value of 4,127,799. The histogram generated by the `skim()` function also indicates a right-skewed distribution for this attribute. This skewness suggests that the majority of borrowers have a moderate amount of money in their accounts, with a smaller proportion having much higher balances. The presence of outliers (see R Code Chunk above) , including values such as 4,127,799, indicates that some borrowers have significantly more money in their accounts, which could have implications for the lending process. Further exploration, such as identifying and possibly removing outliers, could provide more insights into the distribution and potential implications for the lending process.
- The standard deviation (SD) of 151,513.69, from the `skim()`function, indicates the average deviation of individual data points from the mean of 138,467. A standard deviation of approximately 151,513.69 suggests that the distribution of the total current balance of all accounts exhibits a considerable level of variability around the mean. This means that while most borrowers have a moderate total balance clustered around the mean, there is also substantial variability, with a significant portion of them having considerably lower or higher total balances. This variability highlights the diverse financial situations among borrowers, underscoring the importance of considering such differences in the lending process.
- There are 3,427 missing values, which is insignificant , as also the complete_rate has a relatively high value of 92%, indicating completeness in the data.
- In the context of the sector, this could indicate that borrowers with lower current total balances are more inclined to apply for loans than those with higher balances. This is attributed to the fact that individuals with lower balances have significantly less equity available to them, thus potentially necessitating a greater reliance on loans to meet their financial needs.

Task 3. Univariate Exploratory Data Analysis (EDA) with ggplot2

In the following, a univariate exploratory data analysis (EDA) was carried out with the 'myLCdata' data set, which comprises five attributes from Table 1. Therefore a plot was created for each attribute.

Employment Length

```
# Sort the categories
sort_emp_length <- factor(myLCdata$emp_length, levels = c("< 1 year", "1 year", "2 years", "3 years", "4

# Plotting histogram for Employment Length
ggplot(data = myLCdata, aes(x = sort_emp_length)) +
  geom_bar(fill = "skyblue", color = "black") +
  scale_y_continuous(breaks = seq(0, 15000, by = 2500), limits = c(0, 15400),
                     labels = scales::comma_format()) + # format y-axis
  geom_text(stat='count', aes(label=after_stat(count)), vjust=-0.5) +  # Add count labels above bars
  labs(title = "Distribution of Employment Length",
       x = "Employment Length (years)",
       y = "Count") +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1)) # Rotate x-axis labels
```

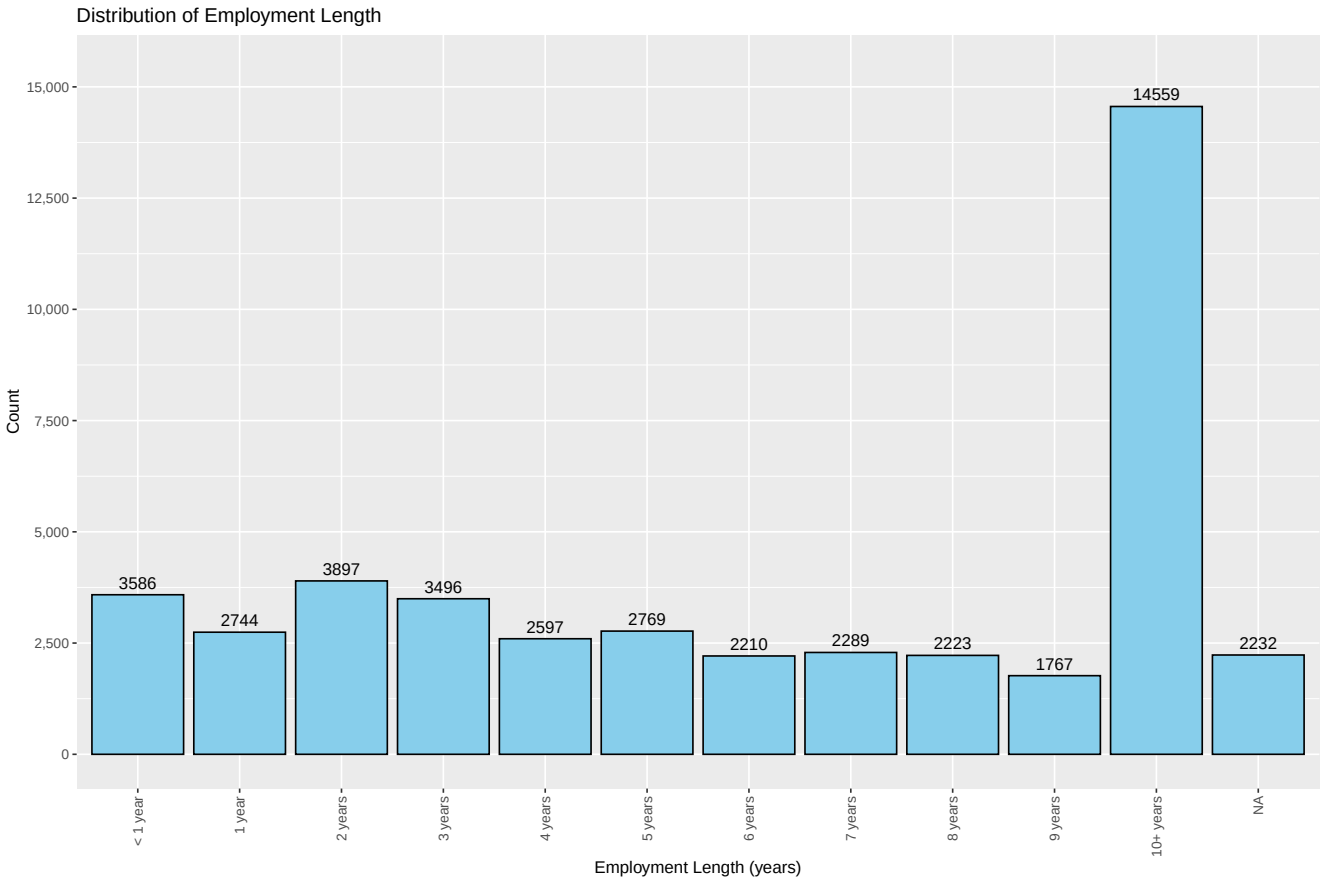

Figure 1: Histogram of Distribution of Employment Length

The histogram depicted in Figure 1 illustrates the distribution of employment length categories on the x-axis, sorted from "<1 year" to "10+ years" and NAs. The y-axis represents the count of loan records associated with each employment length category or NA, with count labels displayed above the bars. The counts vary across categories, ranging from a minimum of 1,767 in "9 years" to a maximum of 14,559 in "10+ years".

The histogram in Figure 1 depicts the same pattern as described in Table 3, revealing a notable disparity in counts among the categories, indicating an imbalance in the distribution of employment lengths. Specifically, a larger proportion of borrowers exhibit longer employment lengths (10+ years: 14,559 occurrences, 32.81% of all occurrences) compared to those with shorter employment lengths (e.g., $<$ 1 year: 3,586 occurrences, 8.08% of all occurrences, or 1 year: 2,744 occurrences, 6.18% of all occurrences). The observation from the histogram in Figure 1 aligns with my assumptions in Section Task 2.

In the domain context, a longer period of employment could indicate career advancement and a higher income. Additionally, individuals with longer employment periods are likely to be older and have accumulated more experience in the job market. This may signal greater financial stability and reliability, potentially reducing the perceived credit risk for lenders. Similarly, longer periods of employment could indicate a stronger credit profile, which could increase the likelihood of lending to these individuals.

Months Since Last Delinquency

```r
# Calculate mean and median with NA values excluded
mean_mths_since_last_delinq <- round(mean(myLCdata$mths_since_last_delinq, na.rm = TRUE), 2)
median_mths_since_last_delinq <- round(median(myLCdata$mths_since_last_delinq, na.rm = TRUE), 2)

# Define bin and bin breaks
bin_width = 4
bin_breaks <- c(seq(0, 144, by = 4), Inf)  # Define bin breaks with exclusive upper bounds

# Group data into bins and calculate counts per bin
bins <- cut(myLCdata$mths_since_last_delinq, na.rm = TRUE, breaks = bin_breaks, include.lowest = TRUE, r
bin_counts <- table(bins)  # Calculate counts per bin
# max(bin_counts) # = 1649 to check the range on the y axis, range of x - axis is the min = 0 and max =

# Calculate the center of each bin for text placement
bin_centers <- bin_breaks[-1] - bin_width / 2

# Create a histogram with mean and median lines
ggplot(data = myLCdata, aes(x = mths_since_last_delinq, na.rm = TRUE)) +
  geom_histogram(binwidth = bin_width, fill = "lightblue", color = "black", breaks = bin_breaks, closed
  geom_vline(aes(xintercept = mean_mths_since_last_delinq, color = "Mean"), linetype = "dashed", linewic
  geom_vline(aes(xintercept = median_mths_since_last_delinq, color = "Median"), linetype = "dotted", lin
  geom_text(data = as.data.frame(bin_counts), aes(label = bin_counts, x = bin_centers, y = bin_counts),
  scale_x_continuous(breaks = seq(0, (max(myLCdata$mths_since_last_delinq, na.rm = TRUE) + bin_width + 6
  scale_y_continuous(breaks = seq(0, 2000, by = 250), limits = c(0, 1750), labels = scales::comma_format
  labs(title = "Distribution of Months Since Last Delinquency",
       x = "Months Since Last Delinquency (months)",
       y = "BiCount") +
  scale_color_manual(name = "Lines", values = c("Mean" = "red", "Median" = "blue"),
                     labels = c(paste("Mean =", mean_mths_since_last_delinq), paste("Median =", median_m
```

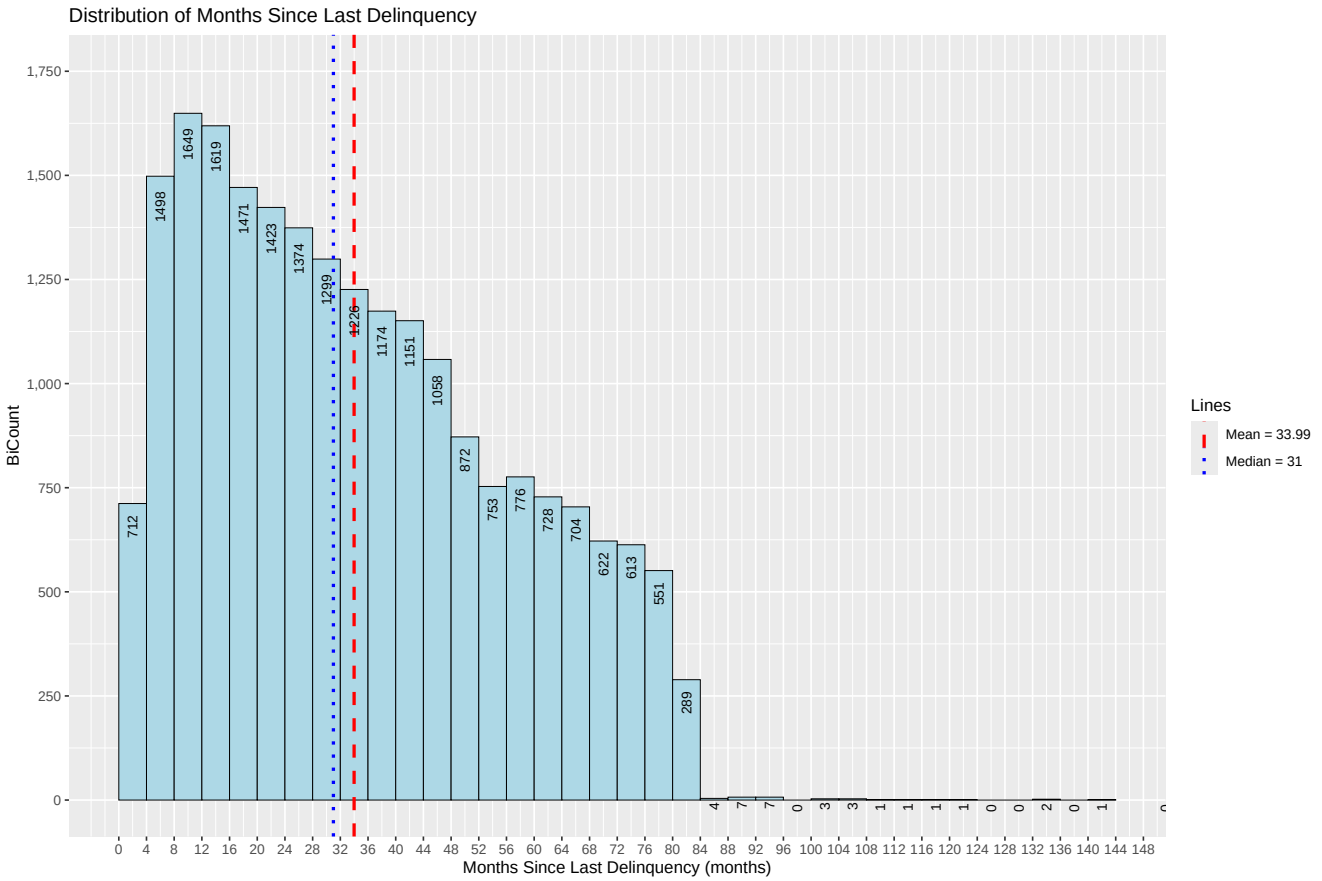

Figure 2: Histogram of Distribution of Months since the last Delinquency

The histogram in Figure 2 illustrates the distribution of months since the last delinquency. The x-axis represents the time in months since the last delinquency, sorted in increasing order, with bins of width 4 used to group the data. The y-axis displays the count of loan records per bin, with count labels presented within the bars. The count per bin ranges from 0 to 1,649, indicating the number of loan records within each bin. The range of months since the last delinquency spans from 0 to 143, as obtained from Section Task 2 min and max value. The NA values were removed and therefore not showing in the histogram. Additionally, the mean and median are highlighted in red and blue, respectively.

The histogram in Figure 2 depicts a right-skewed distribution of the data, suggesting that most borrowers have relatively shorter periods since their last delinquency, while a smaller proportion experience longer periods. The presence of outliers, notably the maximum value of 143 months, indicates that some borrowers even have experienced significantly longer periods since their last delinquency. This aligns with what was interpreted in Section Task 2.

In the domain context it could mean that individuals with short periods since their last delinquency may have turned to alternative sources of credit, such as this peer-to-peer lending platform, as this may have less stringent credit requirements than traditional lenders, such as banks.

```r
# to uncomment this code highlight it an press Ctrl + Shift + C.
# # Calculate counts for each category of mths_since_last_delinq, including NAs
# mths_since_last_delinq_counts <- table(myLCdata$mths_since_last_delinq, useNA = "always")
#
# # Sort the categories
# sorted_mths_since_last_delinq <- c(sort(unique(myLCdata$mths_since_last_delinq)), NA)
#
# # Reorder the counts based on the sorted categories
# mths_since_last_delinq_counts <- mths_since_last_delinq_counts[match(sorted_mths_since_last_delinq, na
#
# # Combine counts and percentages into a data frame
# mths_since_last_delinq_summary <- data.frame(
#   mths_since_last_delinq = names(mths_since_last_delinq_counts),
#   count = as.vector(mths_since_last_delinq_counts))
#
# # Print the table using kable
# knitr::kable(mths_since_last_delinq_summary)
```

This R Code Chunk would generate a table with the counts for each category of "mths_since_last_delinq", including the NA values. The table is helpful to see the distribution within the bins (binwidth = 4) and helped me to check the correctness of the count label, displayed in the bars. However, I have commented it out as it increases the size of the output file, making it less manageable.

Open Credit Lines

```r
# Calculate mean and median with NA values excluded
mean_open_acc <- round(mean(myLCdata$open_acc, na.rm = TRUE), 2)
median_open_acc <- round(median(myLCdata$open_acc, na.rm = TRUE), 2)

# Define bin and bin breaks
bin_width = 2
bin_breaks <- c(seq(0, 77, by = 2), Inf)  # Define bin breaks with exclusive upper bounds

# Group data into bins and calculate counts per bin
bins <- cut(myLCdata$open_acc, na.rm = TRUE, breaks = bin_breaks, include.lowest = TRUE, right = FALSE)
bin_counts <- table(bins)  # Calculate counts per bin
# max(bin_counts) # = 7879 to check the range on the y axis, range of x - axis is the min = 0 and max =

# Calculate the center of each bin for text placement
bin_centers <- bin_breaks[-1] - bin_width / 2

# Create a histogram with mean and median lines
ggplot(data = myLCdata, aes(x = open_acc, na.rm = TRUE)) +
  geom_histogram(binwidth = bin_width, fill = "lightblue", color = "black", breaks = bin_breaks, closed
  geom_vline(aes(xintercept = mean_open_acc, color = "Mean"), linetype = "dashed", linewidth = 1) +
  geom_vline(aes(xintercept = median_open_acc, color = "Median"), linetype = "dotted", linewidth = 1) +
  geom_text(data = as.data.frame(bin_counts), aes(label = bin_counts, x = bin_centers, y = bin_counts),
  scale_x_continuous(breaks = seq(0, (max(myLCdata$open_acc, na.rm = TRUE) + bin_width + 2), by = bin_wi
  scale_y_continuous(breaks = seq(0, 9000, by = 500), limits = c(0, 8500), labels = scales::comma_format
  labs(title = "Distribution of Open Credit Lines",
       x = "Number of Open Credit Lines",
       y = "Count") +
  scale_color_manual(name = "Lines", values = c("Mean" = "red", "Median" = "blue"),
                     labels = c(paste("Mean =", mean_open_acc), paste("Median =", median_open_acc)))
```

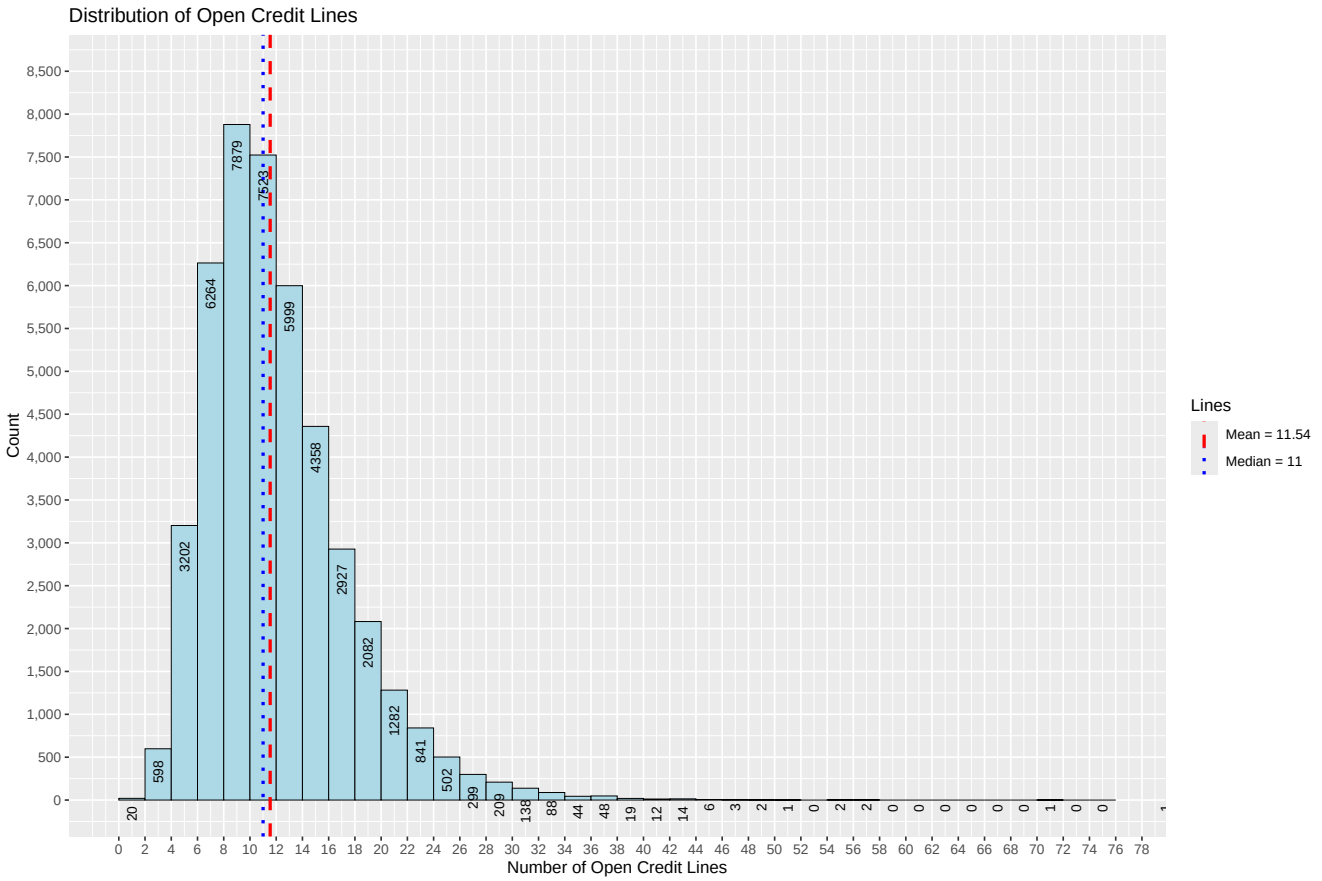

Figure 3: Histogram of Distribution of Open Credit Lines

The histogram in Figure 3 illustrates the distribution of open credit lines. The x-axis represents the number of open credit lines, sorted in increasing order, with bins of width 2 used to group the data. The y-axis displays the count of loan records per bin, with count labels presented within the bars. The count per bin ranges from 0 to 7,879, indicating the number of loan records within each bin. The range of open credit lines spans from 0 to 76, as obtained from Section Task 2 min and max value. The NA values were removed and therefore not showing in the histogram. Additionally, the mean and median are highlighted in red and blue, respectively.

The histogram in Figure 3 depicts a right-skewed distribution of the data, suggesting that the majority of borrowers have relatively moderate amount of open credit lines, with a smaller proportion that has many more open credit lines. The presence of outliers, notably the maximum value of 76 open credit lines, indicates that some borrowers even have significantly more open credit lines. This aligns with what was interpreted in Section Task 2.

In the domain context, this could suggest that lenders perceive borrowers with a low to moderate number of open credit lines as more stable and financially responsible. Consequently, such borrowers would be more appealing candidates for obtaining a loan.

```
# to uncomment this code highlight it an press Ctrl + Shift + C.
# # Calculate counts for each category of open_acc, including NAs
# open_acc_counts <- table(myLCdata$open_acc, useNA = "always")
#
# # Sort the categories
# sorted_open_acc <- c(sort(unique(myLCdata$open_acc)), NA)
#
# # Reorder the counts based on the sorted categories
# open_acc_counts <- open_acc_counts[match(sorted_open_acc, names(open_acc_counts))]
#
# # Combine counts and percentages into a data frame
# open_acc_summary <- data.frame(
#   open_acc = names(open_acc_counts),
#   count = as.vector(open_acc_counts))
#
# # Print the table using kable
# knitr::kable(open_acc_summary)
```

This R Code Chunk would generate a table with the counts for each category of "open_acc", including the NA values. The table is helpful to see the distribution within the bins (binwidth = 2) and helped me to check the correctness of the count label, displayed in the bars. However, I have commented it out as it increases the size of the output file, making it less manageable.

Purpose of Loan

```
# Sort purposes by frequency in increasing order
sorted_purpose <- names(sort(table(myLCdata$purpose)))

# Plotting bar plot for Purpose of Loan with sorted labels and counts as labels
ggplot(data = myLCdata, aes(x = factor(purpose, levels = sorted_purpose))) +
  geom_bar(fill = "skyblue", color = "black") +
  scale_y_continuous(breaks = seq(0, 30000, by = 5000), limits = c(0, 27000), labels = scales::comma_for
  geom_text(stat='count', aes(label=after_stat(count)), vjust=-0.5) +  # Add count labels above bars
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  labs(title = "Distribution of Loan Purposes",
       x = "Purpose of Loan",
       y = "Count")
```

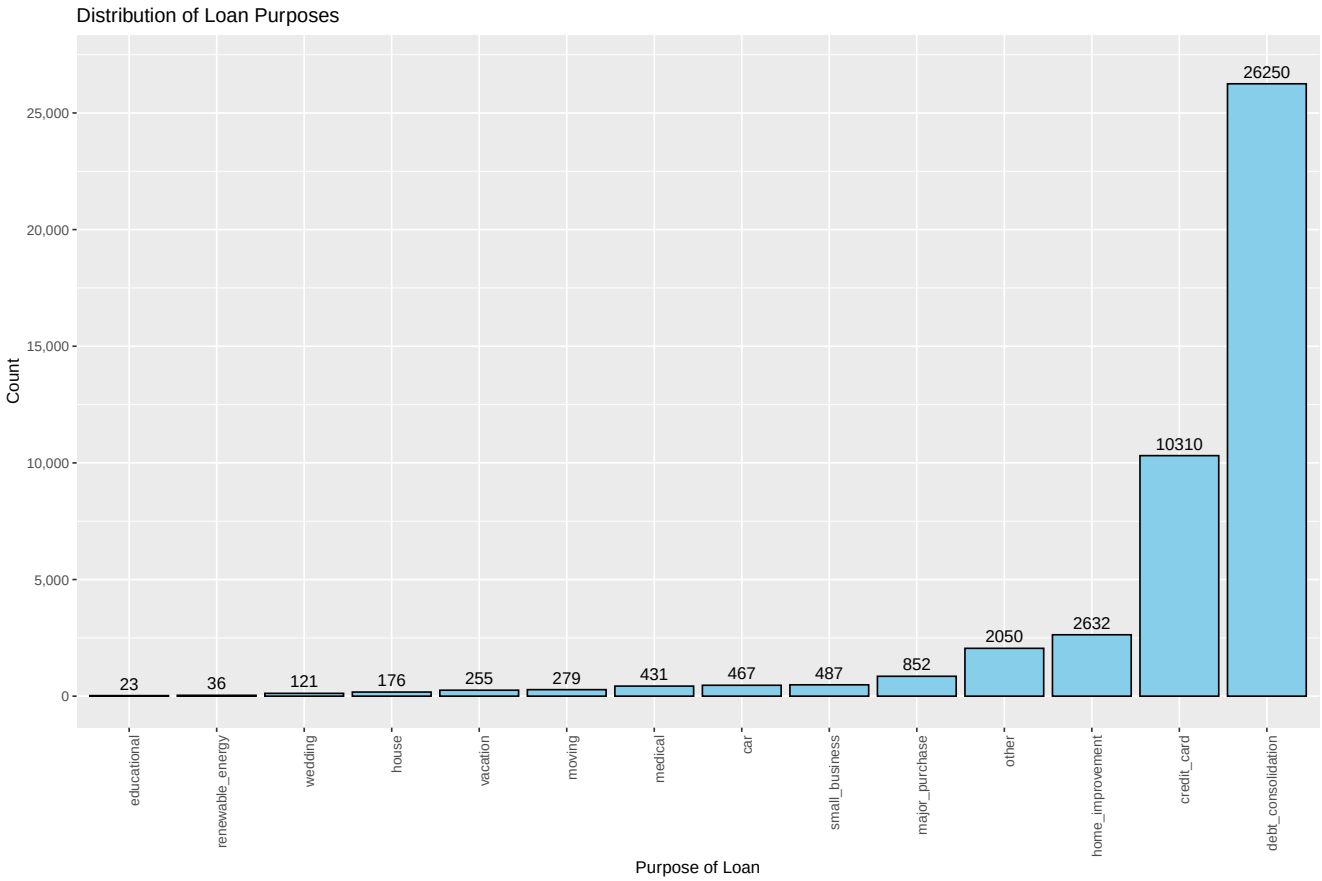

Figure 4: Histogram of Distribution of Loan Purposes

The histogram depicted in Figure 4 illustrates the distribution of loan purpose categories on the x-axis, sorted in increasing order according to the count of loan records per category. The y-axis represents the count of loan records associated with each loan purpose category, with count labels displayed above the bars. The counts vary across categories, ranging from a minimum of 23 for "educational" to a maximum of 26,250 for "debt consolidation". As observed in Section Task 2, there are no NAs recorded for this attribute.

The histogram presented in Figure 4 showcases a pattern consistent with the findings outlined in @tbl-frequencies-purpose, wherein "debt_consolidation" and "credit_card" dominate the dataset. These categories typically signify financial obligations or debt repayment, which may be perceived as less favorable. In contrast, less common categories such as "educational" and "renewable_energy" imply investments in education or environmentally friendly initiatives, potentially leading to enhanced financial stability or savings. The notable disparity between these categories underscores an imbalance within the dataset, emphasizing the prevalence of loan applications aimed at debt management compared to those focused on self-improvement or sustainable practices. The observation from the histogram in @fig-histogram_emp_length aligns with my assumptions in Section Task 2.

In the domain context, the prevalence of loan applications for purposes such as "debt_consolidation" and "credit_card" suggests that individuals utilizing this platform may be facing financial challenges or existing debt obligations. This pattern could imply that traditional financial institutions, like banks, may not perceive these individuals as creditworthy or trustworthy enough to extend further credit, leading them to seek alternative borrowing options. Consequently, they may resort to platforms like this one to consolidate their debts or manage existing financial burdens.

Total Current Balance

```
# Calculate mean and median with NA values excluded
mean_tot_cur_bal <- round(mean(myLCdata$tot_cur_bal, na.rm = TRUE), 1)
median_tot_cur_bal <- round(median(myLCdata$tot_cur_bal, na.rm = TRUE), 1)

# Define bin and bin breaks
bin_width = 100000
bin_breaks <- c(seq(0, 4127800, by = 100000), Inf)  # Define bin breaks with exclusive upper bounds

# Group data into bins and calculate counts per bin
bins <- cut(myLCdata$tot_cur_bal, na.rm = TRUE, breaks = bin_breaks, include.lowest = TRUE, right = FALS
bin_counts <- table(bins)  # Calculate counts per bin
# max(bin_counts) # = 22336 to check the range on the y axis, range of x - axis is the min = 0 and max =

# Calculate the center of each bin for text placement
bin_centers <- bin_breaks[-1] - bin_width / 2

# Create a histogram with mean and median lines
ggplot(data = myLCdata, aes(x = tot_cur_bal, na.rm = TRUE)) +
  geom_histogram(binwidth = bin_width, fill = "lightblue", color = "black", breaks = bin_breaks, closed
  geom_vline(aes(xintercept = mean_tot_cur_bal, color = "Mean"), linetype = "dashed", linewidth = 1) +
  geom_vline(aes(xintercept = median_tot_cur_bal, color = "Median"), linetype = "dotted", linewidth = 1)
  geom_text(data = as.data.frame(bin_counts), aes(label = bin_counts, x = bin_centers, y = bin_counts),
  scale_x_continuous(breaks = seq(0, (max(myLCdata$tot_cur_bal, na.rm = TRUE) + bin_width), by = bin_wid
labels = scales::unit_format(unit = "k", scale = 1e-3, sep = "")) +  # Set breaks and adjust limits
  scale_y_continuous(breaks = seq(0, 22500, by = 500), limits = c(0, 23000),
labels = scales::comma_format()) + # Set breaks every 500 and format with comma separator
  labs(title = "Distribution of Total Current Balance",
       x = "Total Current Balance in $k",
       y = "Count") +
  scale_color_manual(name = "Lines", values = c("Mean" = "red", "Median" = "blue"),
                     labels = c(paste("Mean =", mean_tot_cur_bal), paste("Median =", median_tot_cur_bal)
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1))
```

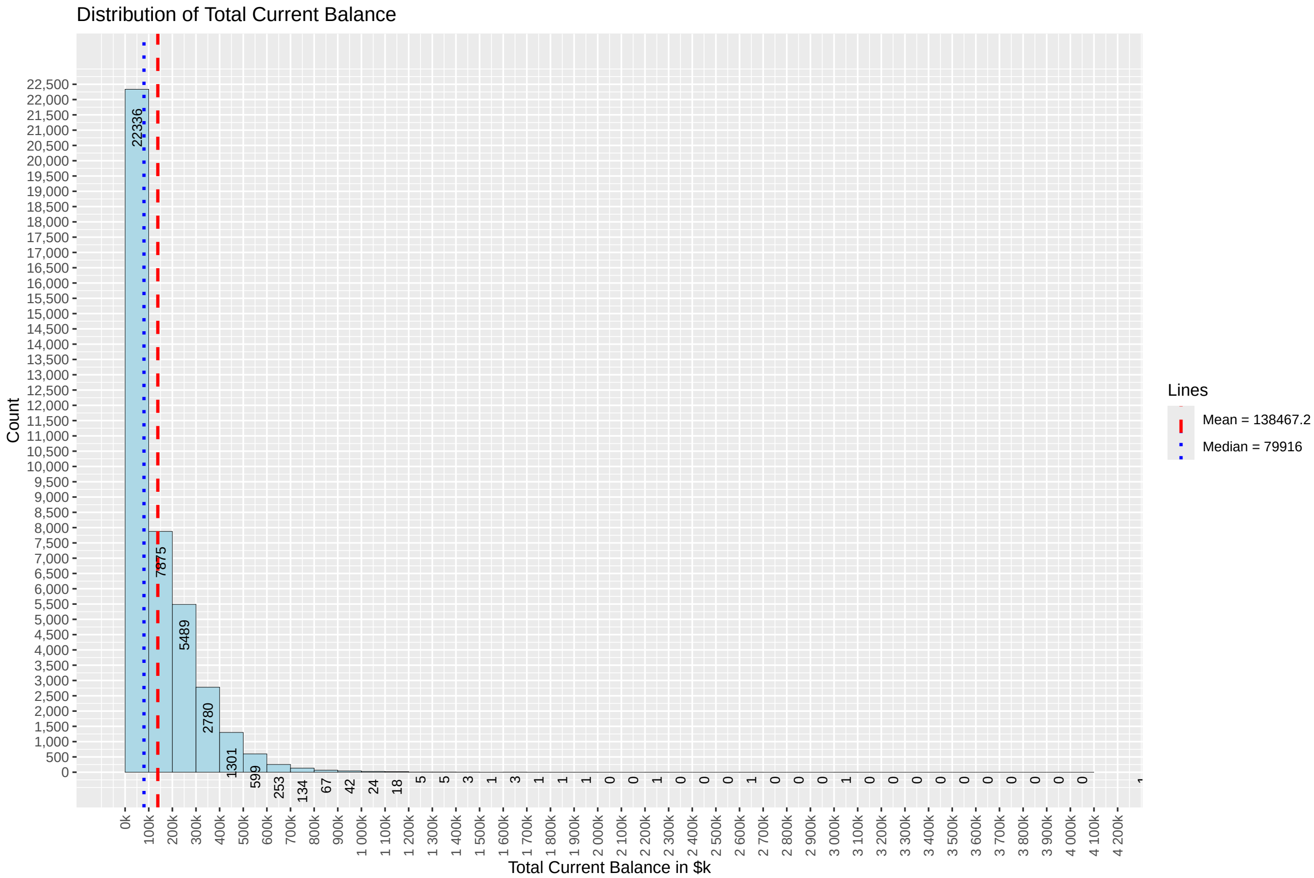

Figure 5: Histogram of Distribution of Total Current Balances

The histogram in Figure 5 illustrates the distribution of total current balance. The x-axis represents the current balance, sorted in increasing order, with bins of width 100,000 used to group the data. The y-axis displays the count of loan records per bin, with count labels presented within the bars. The count per bin ranges from 0 to 22,336, indicating the number of loan records within each bin. The range of the total current balance s spans from 0 to 4,127,799, as obtained from Section Task 2 min and max value. The NA values were removed and therefore not showing in the histogram. Additionally, the mean and median are highlighted in red and blue, respectively.

The histogram shown in Figure 5 displays a right-skewed distribution of the data, indicating that the majority of borrowers maintain relatively lower total balances. However, a small proportion of borrowers exhibit substantially higher balances, with some individuals having significantly more money in their accounts. This aligns with what was interpreted in Section Task 2.

In the context of the sector, this could indicate that borrowers with lower current total balances are more inclined to apply for loans than those with higher balances. This is attributed to the fact that individuals with lower balances have significantly less equity available to them, thus potentially necessitating a greater reliance on loans to meet their financial needs.

```
# to uncomment this code highlight it an press Ctrl + Shift + C.
# # Calculate counts for each category of tot_cur_bal, including NAs
# tot_cur_bal_counts <- table(myLCdata$tot_cur_bal, useNA = "always")
#
# # Sort the categories
# sorted_tot_cur_bal <- c(sort(unique(myLCdata$tot_cur_bal)), NA)
#
# # Reorder the counts based on the sorted categories
# tot_cur_balcounts <- tot_cur_bal_counts[match(sorted_open_acc, names(tot_cur_bal_counts))]
#
# # Combine counts and percentages into a data frame
# tot_cur_bal_summary <- data.frame(
#   tot_cur_bal = names(tot_cur_bal_counts),
#   count = as.vector(tot_cur_bal_counts))
#
# # Print the table using kable
# knitr::kable(tot_cur_bal_summary)
```

This R Code Chunk would generate a table with the counts for each category of "total_cur_bal", including the NA values. The table is helpful to see the distribution within the bins (binwidth = 100,000) and helped me to check the correctness of the count label, displayed in the bars. However, I have commented it out as it significantly increases the size of the output file, making it less manageable.

Task 4. Bivariate EDA with ggplot2

```
# Create pairs plot
pairs_plot <- ggpairs(myLCdata)
pairs_plot
```

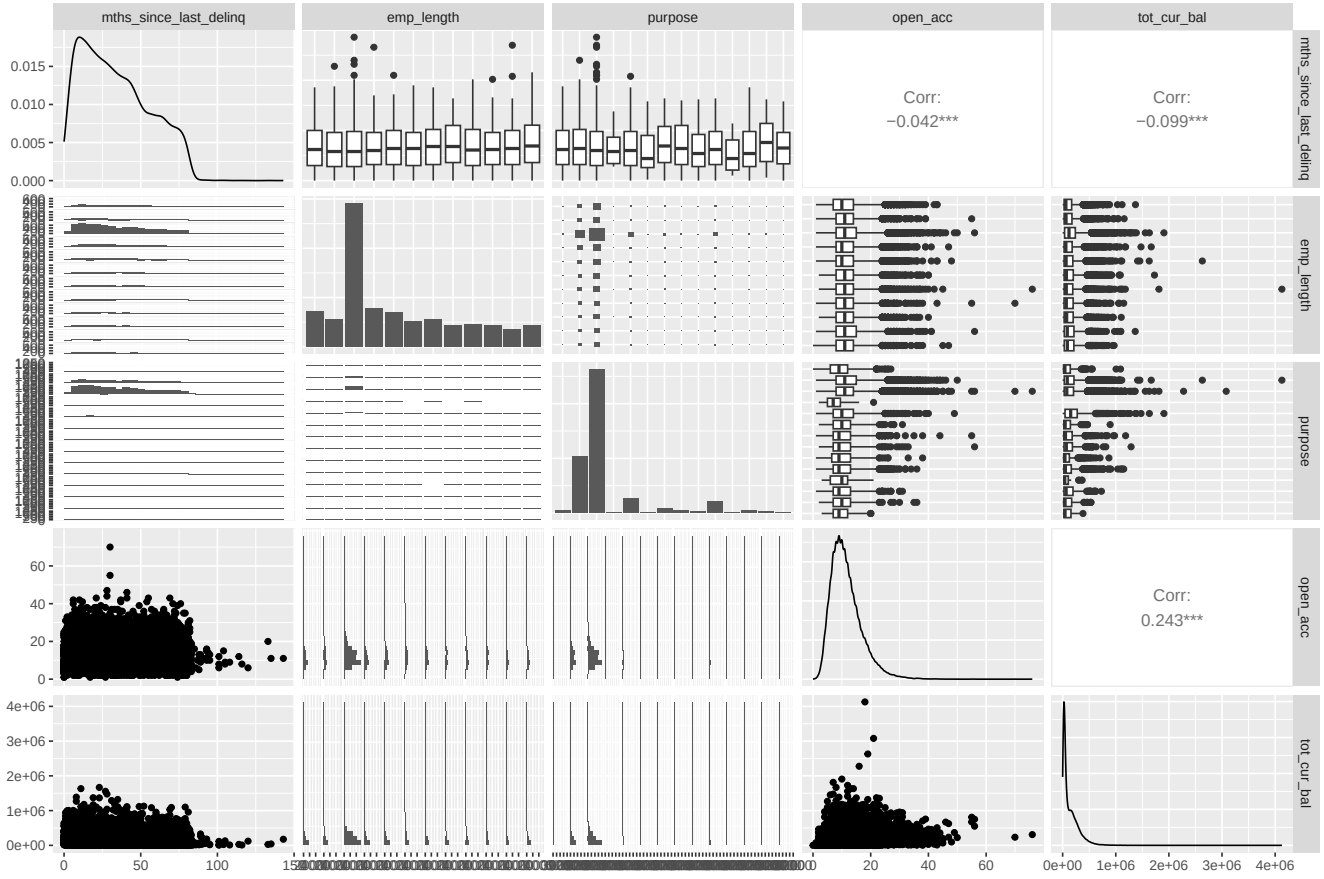

Figure 6: Pairs plot of the data to get a first impression.

The pairs plot depicted in Figure 6 provides a comprehensive visualization of pairwise relationships and distributions within the data set. Each cell in the grid represents a plot of two attributes, allowing for visual examination of correlations, distributions, and potential outliers. The diagonal displays density plots or histograms of individual variables, offering insights into their distributions. This plot serves as a valuable tool for gaining a first impression and understanding the interrelationships among variables in the data set. The NA values get automatically removed by the `ggpairs()`function.

2-dimensional Subplots of Pairsplot

Open Credit Lines & Total Current Balance

```
# Calculate the correlation coefficient between open_acc and tot_cur_bal
correlation <- cor(myLCdata$open_acc, myLCdata$tot_cur_bal, use = "complete.obs")

# Create scatterplot with mean and median lines, correlation coefficient, and trendline
scatterplot_open_acc_tot_cur_bal <- ggplot(myLCdata, aes(x = open_acc, y = tot_cur_bal)) +
  geom_point() +
  scale_x_continuous(breaks = seq(0, 80, by = 2), limits = c(0, 78)) +
  scale_y_continuous(breaks = seq(0, 4200000, by = 100000), labels = scales::unit_format(unit = "k", sca
  labs(x = "Open Credit Lines", y = "Total Current Balance in $k") +
  annotate("text", x = 70, y = 2300000, label = paste("Correlation:", round(correlation, 2)), color = "b
  geom_smooth(method = "lm", se = FALSE, aes(color = "Trendline")) +  # Add linear trendline without cor
  scale_color_manual(name = "Line", values = c("Trendline" = "purple"))  # Set color for the trendline

scatterplot_open_acc_tot_cur_bal
```

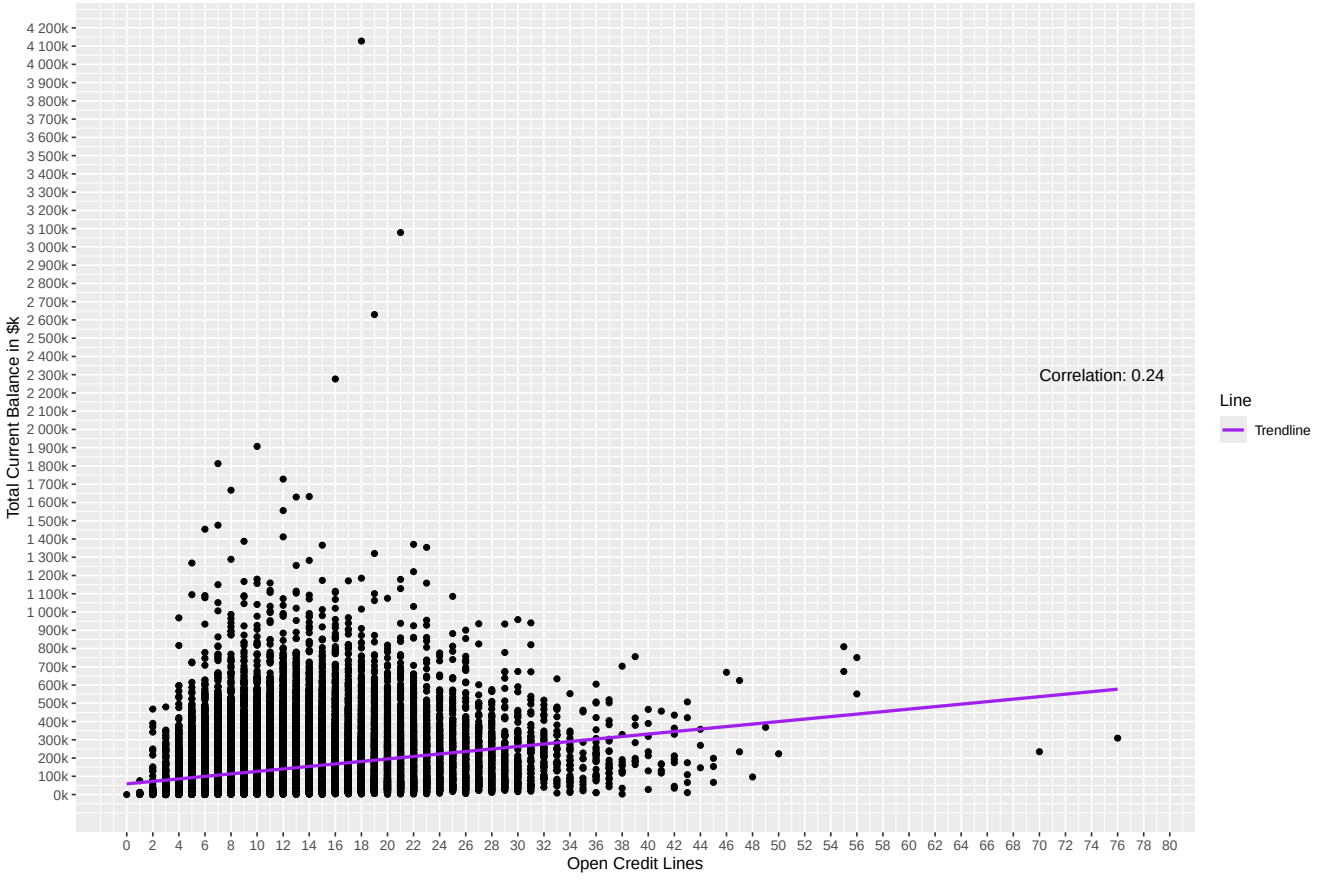

Figure 7: Scatterplot showing the relationship between the open credit lines and the total current balance.

```r
# Create scatterplot with mean and median lines, correlation coefficient, and trendline
scatterplot_open_acc_tot_cur_bal_zoomed <- ggplot(myLCdata, aes(x = open_acc, y = tot_cur_bal)) +
  geom_point() +
  scale_x_continuous(breaks = seq(0, 80, by = 2), limits = c(0, 78)) +
  scale_y_continuous(breaks = seq(0, 1700000, by = 100000), limits = c(0, 1700000), labels = scales::uni
  labs(x = "Open Credit Lines", y = "Total Current Balance in $k") +
  annotate("text", x = 70, y = 1000000, label = paste("Correlation:", round(correlation, 2)), color = "b
  geom_smooth(method = "lm", se = FALSE, aes(color = "Trendline")) +  # Add linear trendline without cor
  scale_color_manual(name = "Line", values = c("Trendline" = "purple"))  # Set color for the trendline

scatterplot_open_acc_tot_cur_bal_zoomed
```

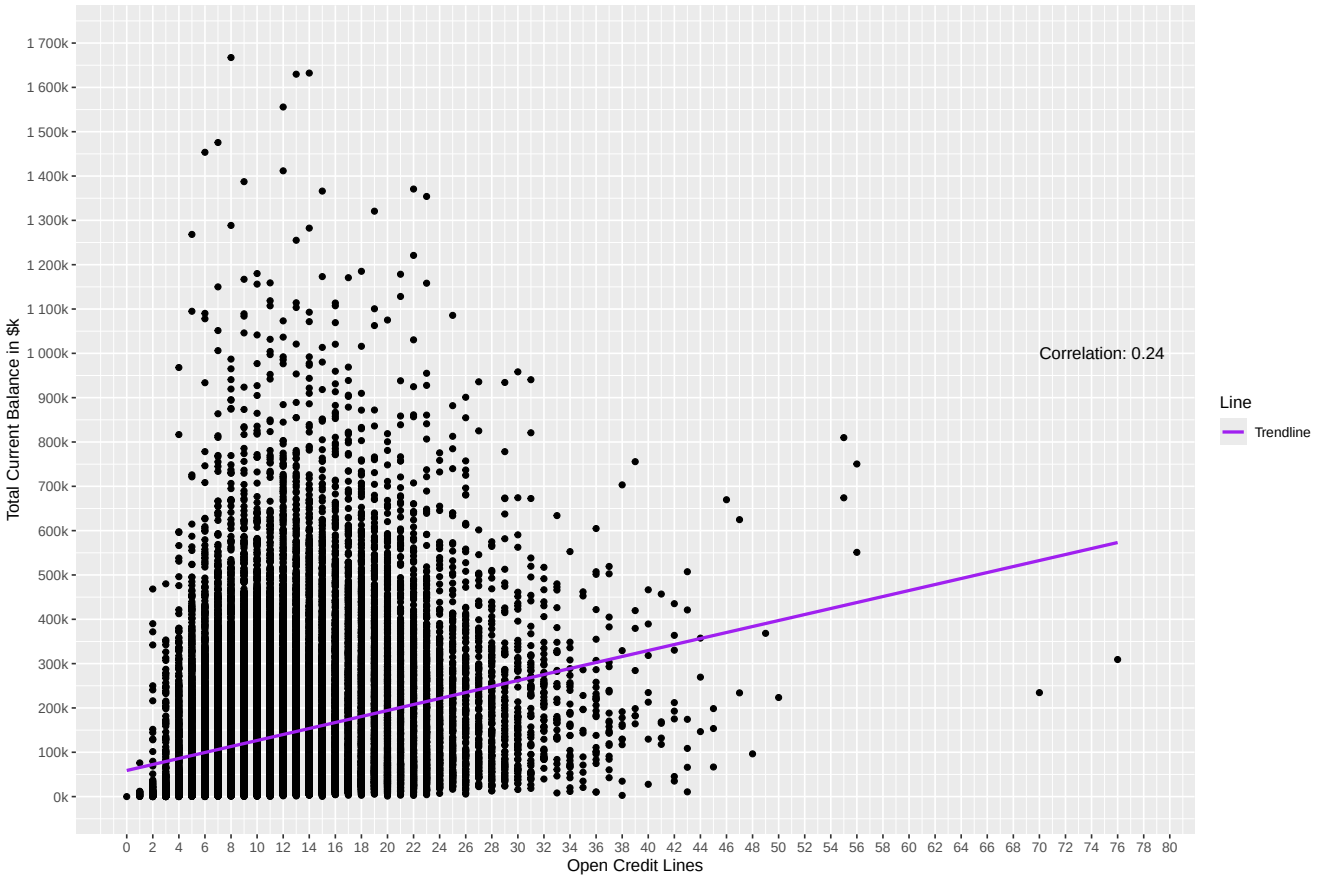

Figure 8: Zoomed in Scatterplot showing the relationship between the open credit lines and the total current balance.

The combination of open_acc and tot_cur_bal in Figure 7 could indicate whether borrowers with a higher number of open credit lines tend to have larger total balances, which could possibly indicate higher credit utilization or more extensive financial activities. The correlation coefficient of 0.24 indicates a positive linear relationship between the variables open_acc (number of open credit lines) and tot_cur_bal (current total balance), which is also evident in the trend line rising to the right.

As the correlation coefficient of 0.24 is relatively small, this suggests a relatively weak positive correlation between the variables. Although there is a positive correlation between the number of open credit lines and the current total balance, it is not particularly strong. It is also important to note that correlation does not imply causality. Even if there is a positive correlation between open_acc and tot_cur_bal, this does not necessarily mean that more open credit lines result in a higher current total balance or vice versa. This is because other factors could influence the relationship between these variables.

If we look at the "zoomed in" scatterplot in Figure 8, we observe that up to around 10-14 open credit lines, more and more individuals with higher total balances (from $0 to around $900,000) have taken out loans. However, thereafter, as the number of outstanding loans increases, there is a tendency for more borrowers with lower overall balances to apply for further loans. This observation suggests that at a certain point, borrowers with higher total balances may require fewer open credit lines, or other factors may come into play that influence the relationship between open credit lines and total balance.

Purpose & Employment Length

```
# Sort emp_length
myLCdata$emp_length <- factor(myLCdata$emp_length, levels = c("< 1 year", "1 year", "2 years", "3 years'

# Sort purpose by count in increasing order
sorted_purpose <- names(sort(table(myLCdata$purpose)))

# Recreate count plot with sorted emp_length and purpose
countplot <- ggplot(myLCdata, aes(x = factor(purpose, levels = sorted_purpose), y = emp_length)) +
   geom_count(aes(color = after_stat(n), size = after_stat(n))) +
    guides(color = 'legend') +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
   labs(x = "Purpose", y = "Employment Length")

countplot
```

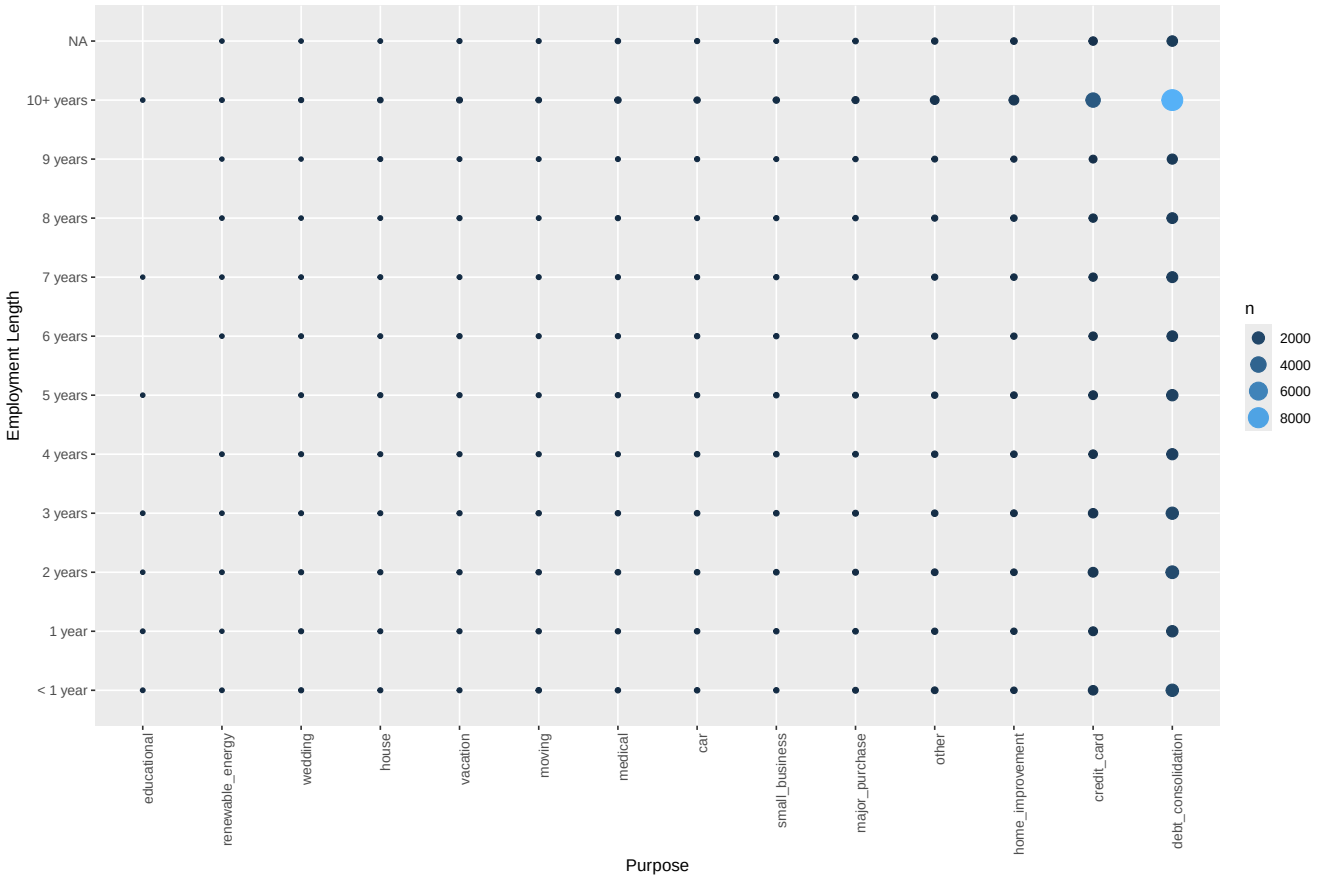

Figure 9: Countplot showing the relationship between the purposes and employment length (n = count).

The relationship between the purpose of a loan and the length of employment (emp_length) in Figure 9 could provide valuable insights into the borrowing behavior of individuals based on their length of employment. Longer employment lengths may indicate greater job stability and financial security. Which could indicate that borrowers with longer employment lengths may be more inclined to apply for loans for purposes such as home construction or education, which typically require a stable financial foundation and long-term planning. Conversely, borrowers with shorter employment histories or unstable employment may be more likely to apply for loans for immediate needs or emergencies, such as debt consolidation or credit card purposes.

However, when we examine Figure 9, we observe that the highest counts of loans, is for for debt consolidation (around 8000) and credit card purposes (around 6000), both from borrowers with 10+ years of employment. (Note: The categories of purpose are sorted according to the count of loan records increasingly per category in the count-plot.) Generally, debt consolidation and credit card purposes, as analyzed in Section Task 3, emerge as the most frequent purposes. Furthermore, the educational loan purpose is predominantly utilized by borrowers with shorter employment durations (<1, 1, 2, 3, 5 years).

We also see in Figure 9 that borrowers with "10+ years" of employment generally have the most loans across all purposes, as already noted in Section Task 3.

Additionally, it should be noted that other factors may have an influence on the purpose of the loan. For example, in times of economic downturn, borrowers may prioritize debt consolidation or paying off credit card bills to overcome financial hardship, regardless of their length of employment.

2-dimensional Subplot of Pairsplot with a 3rd Attribute

As Figure 9 is a countplot of two attributes of type character, a 3rd variable cannot be added to the plot.

Open Credit Lines & Total Current Balance

```
# Create scatterplot with purpose colored
scatterplot_open_acc_tot_cur_bal_purpose <- ggplot(myLCdata, aes(x = open_acc, y = tot_cur_bal, color =
  geom_point() +
  scale_x_continuous(breaks = seq(0, 80, by = 2), limits = c(0, 78)) +
  scale_y_continuous(breaks = seq(0, 4200000, by = 100000), labels = scales::unit_format(unit = "k", sca
  labs(x = "Open Credit Lines", y = "Total Current Balance in $k", color = "Purpose")
scatterplot_open_acc_tot_cur_bal_purpose
```

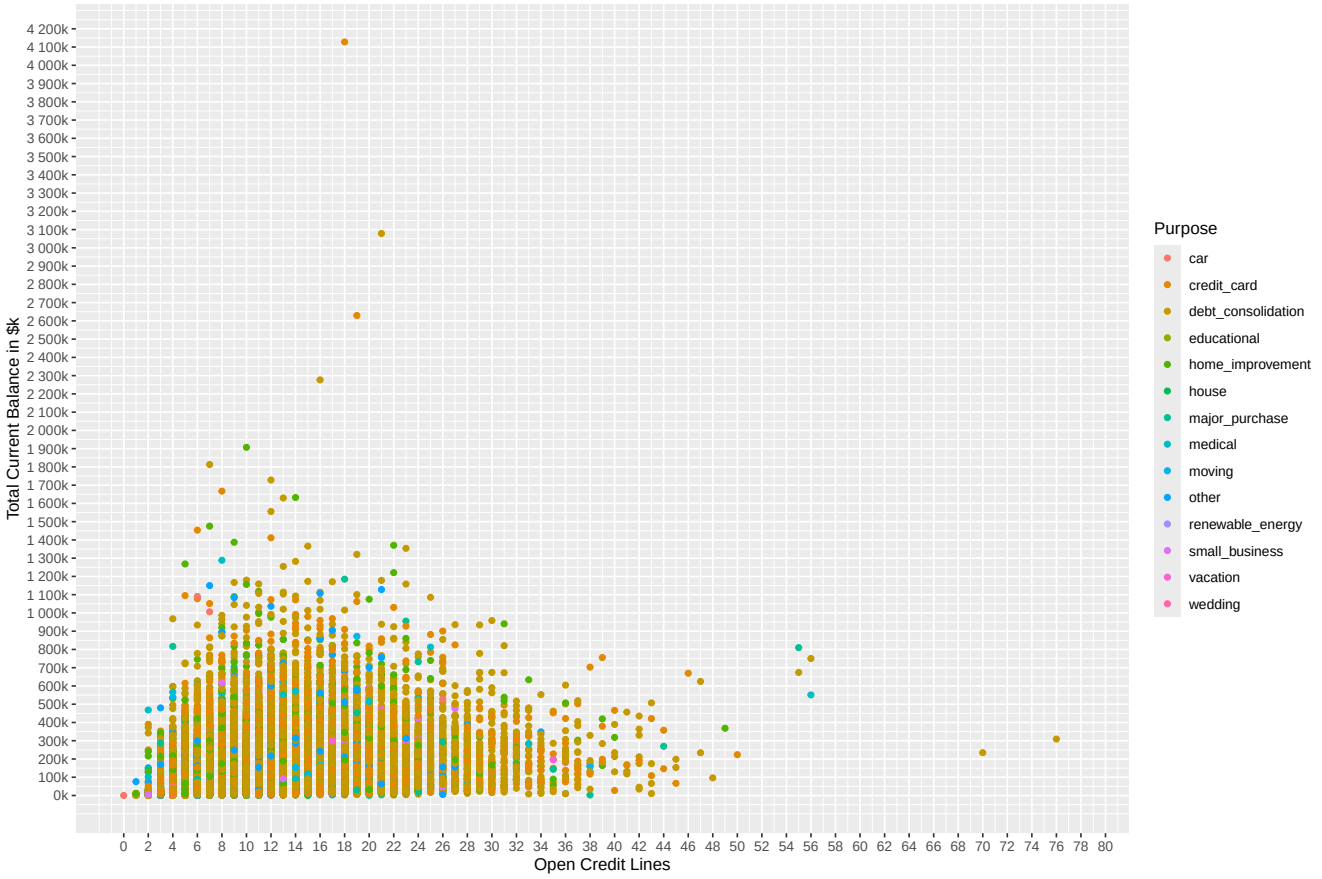

Figure 10: Scatterplot showing the relationship between the open credit lines and the total current balance colored by purpose.

```r
# Create scatterplot with emp_length colored
scatterplot_open_acc_tot_cur_bal_emp_length <- ggplot(myLCdata, aes(x = open_acc, y = tot_cur_bal, color
  geom_point() +
  scale_x_continuous(breaks = seq(0, 80, by = 2), limits = c(0, 78)) +
  scale_y_continuous(breaks = seq(0, 4200000, by = 100000), labels = scales::unit_format(unit = "k", sca
  labs(x = "Open Credit Lines", y = "Total Current Balance in $k", color = "Employment Length")

scatterplot_open_acc_tot_cur_bal_emp_length
```

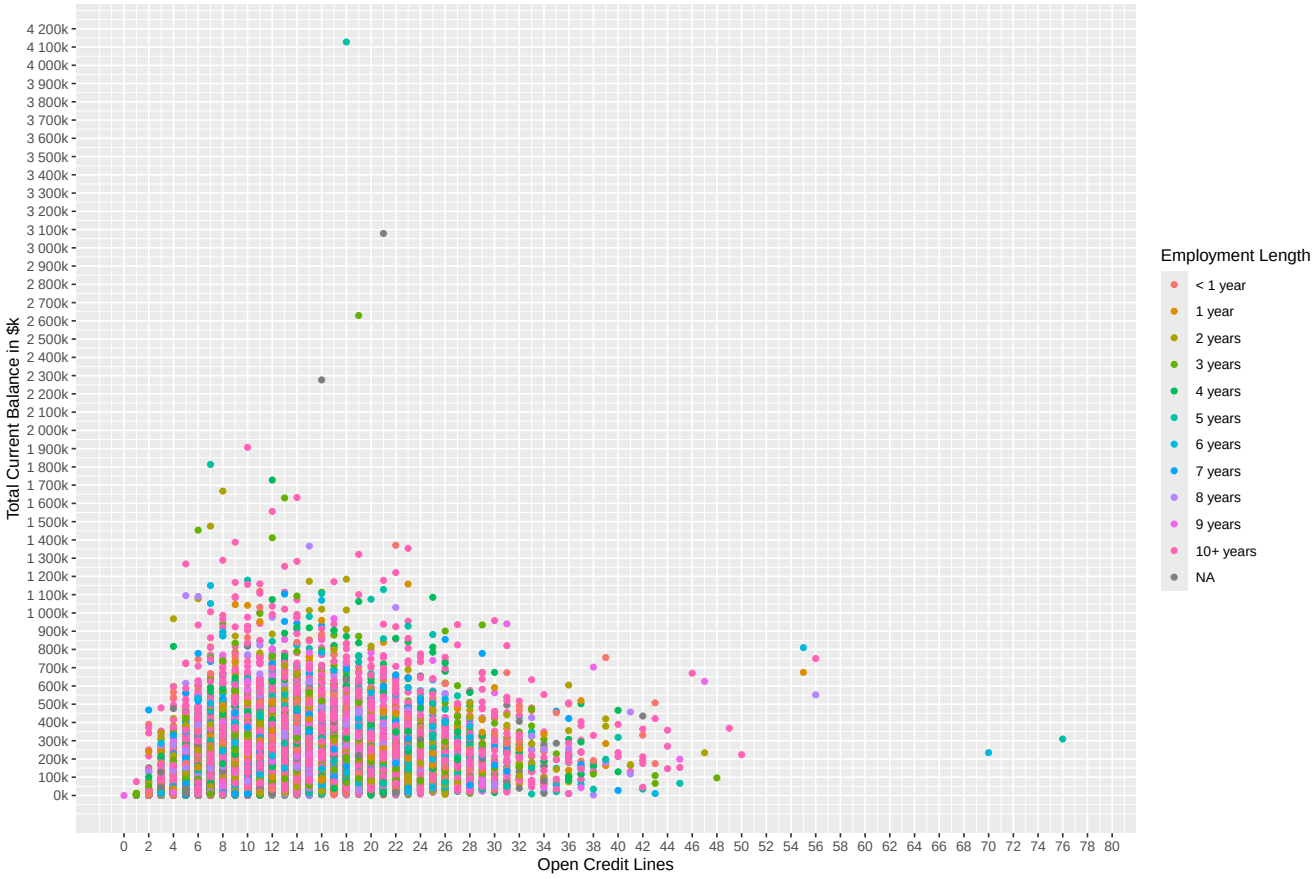

Figure 11: Scatterplot showing the relationship between the open credit lines and the total current balance colored by employment length.

```r
# Define the breaks for the color scale
color_breaks <- seq(0, 160, by = 20)  # Change the step size to 10

# Create scatterplot with mths_since_last_delinq colored
scatterplot_open_acc_tot_cur_bal_mths_since_last_delinq <- ggplot(myLCdata, aes(x = open_acc, y = tot_cu
  geom_point() +
  scale_x_continuous(breaks = seq(0, 80, by = 2), limits = c(0, 78)) +
  scale_y_continuous(breaks = seq(0, 4200000, by = 100000), labels = scales::unit_format(unit = "k", sca
  scale_color_continuous(breaks = color_breaks, low = "#4B0082", high = "#DBB7FF") +  # Set breaks for c
  labs(x = "Open Credit Lines", y = "Total Current Balance in $k", color = "Months Since Last Delinquenc

scatterplot_open_acc_tot_cur_bal_mths_since_last_delinq
```

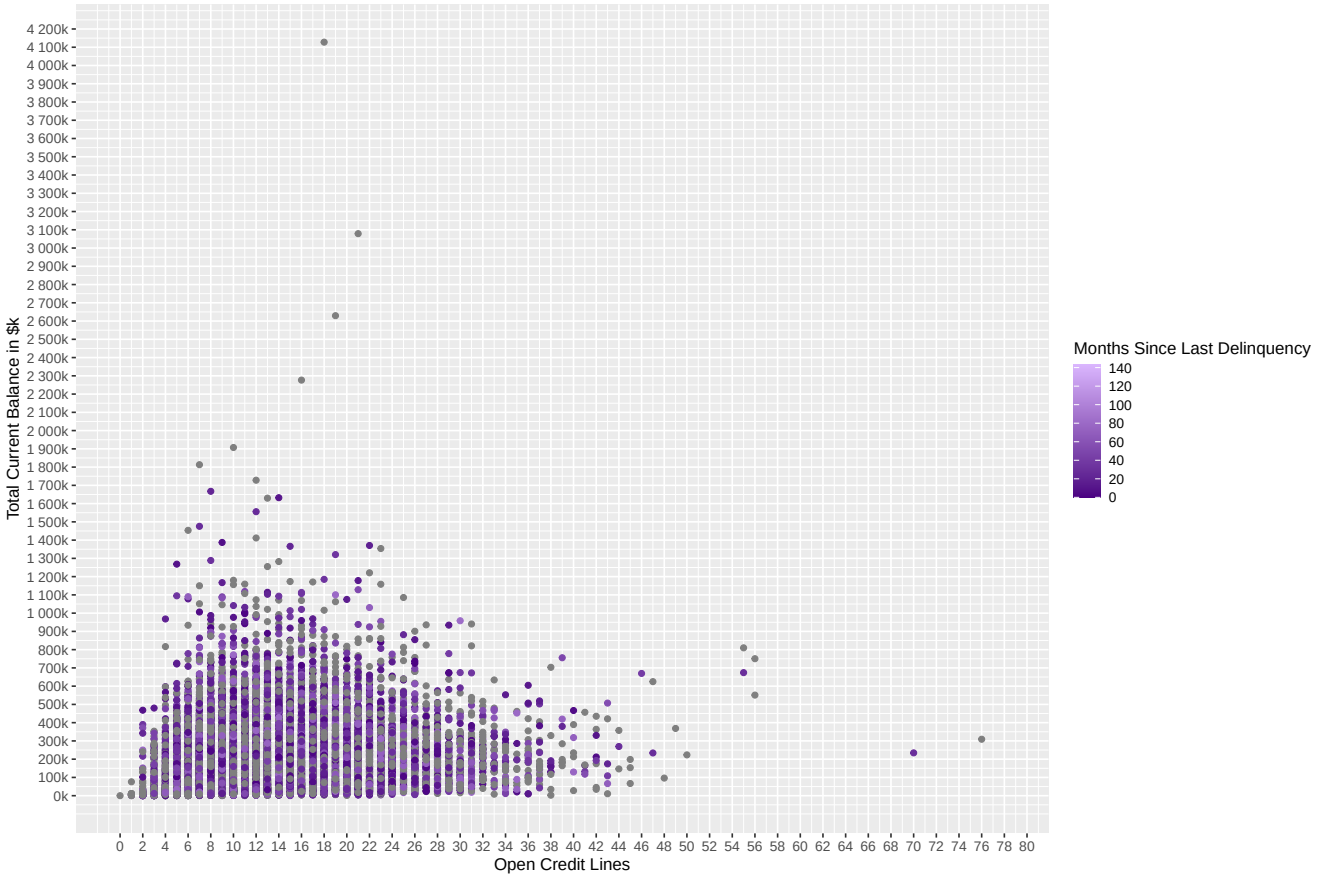

Figure 12: Scatterplot showing the relationship between the open credit lines and the total current balance colored by months since last delinquency.

To get further insights into Figure 7 and its "zoomed in" version in Figure 8 the above plots were created. In Figure 10, the attribute 'purpose' was visualized using color. Similarly, in Figure 11 and Figure 12, 'employment length' and 'months since last delinquency' were represented via color, respectively. Unfortunately, none of these plots revealed any new insights or knowledge, as the added attributes visualized by color did not uncover any discernible clusters, patterns, or trends.

Months Since Last Delinquency & Open Credit Lines

```
# Define breaks for the color scale
color_breaks <- seq(0, 4300000, by = 500000)  # Adjust the step size as needed

# Create scatterplot with mean and median lines, correlation coefficient, and trendline
scatterplot_mths_since_last_delinq_open_acc_tot_cur_bal <- ggplot(myLCdata, aes(x = mths_since_last_deli
  geom_point() +
  scale_x_continuous(breaks = seq(0, 144, by = 4), limits = c(0, 144)) +
  scale_y_continuous(breaks = seq(0, 80, by = 5), labels = scales::unit_format(unit = "k", scale = 1e-3,
  scale_color_continuous(breaks = color_breaks, labels = scales::unit_format(unit = "k", scale = 1e-3, s
  labs(x = "Months Since Last Delinquency", y = "Open Credit Lines", color = "Total Current Balance in $

scatterplot_mths_since_last_delinq_open_acc_tot_cur_bal
```

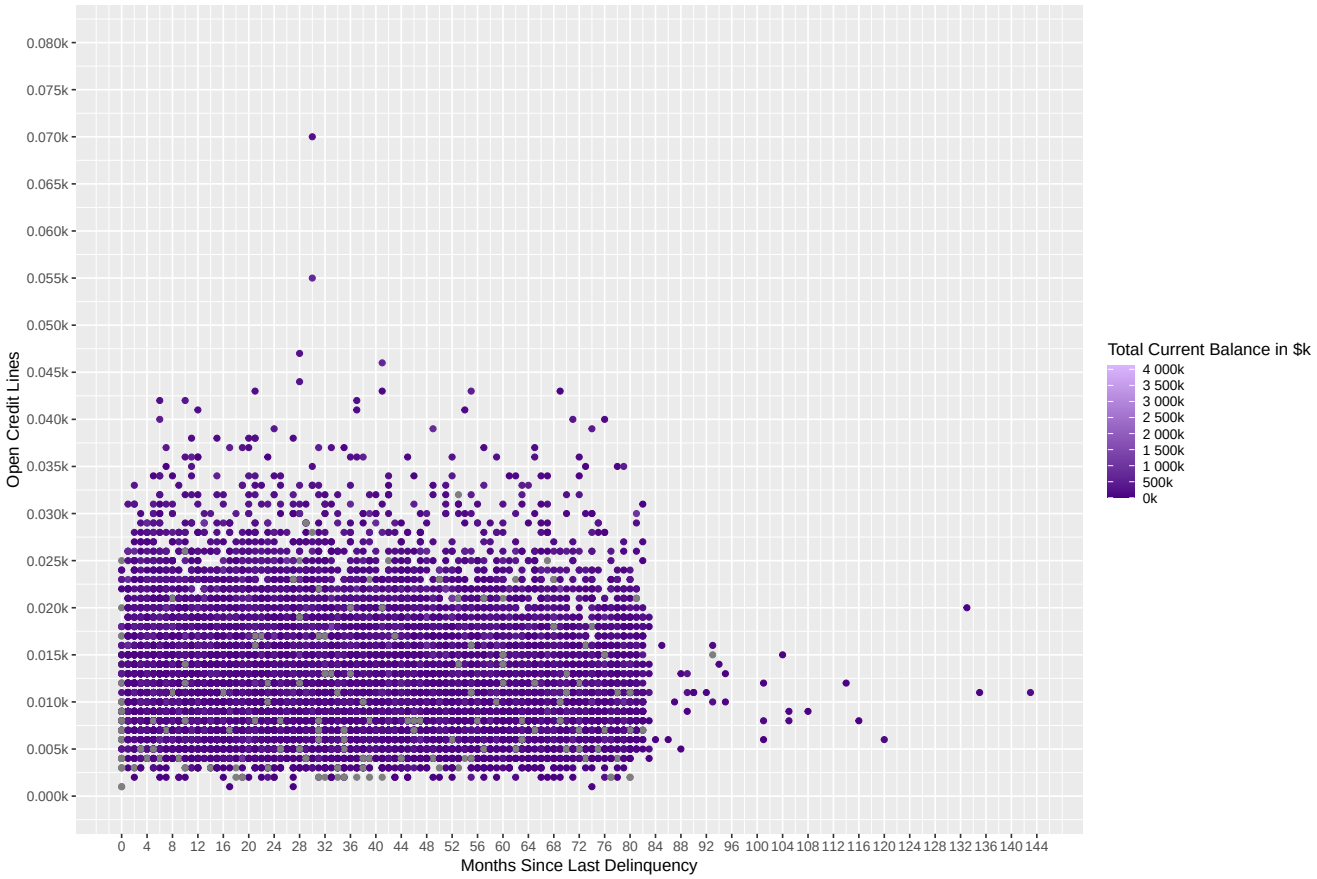

Figure 13: Scatterplot shwoing the relationship between the months since last delinquency and the open credit lines colored by total current balance.

```r
# Create scatterplot with mean and median lines, correlation coefficient, and trendline
scatterplot_mths_since_last_delinq_open_acc_purpose <- ggplot(myLCdata, aes(x = mths_since_last_delinq,
  geom_point() +
  scale_x_continuous(breaks = seq(0, 144, by = 4), limits = c(0, 144)) +
  scale_y_continuous(breaks = seq(0, 80, by = 5), labels = scales::unit_format(unit = "k", scale = 1e-3,
  labs(x = "Months Since Last Delinquency", y = "Open Credit Lines", color = "Purpose")

scatterplot_mths_since_last_delinq_open_acc_purpose
```

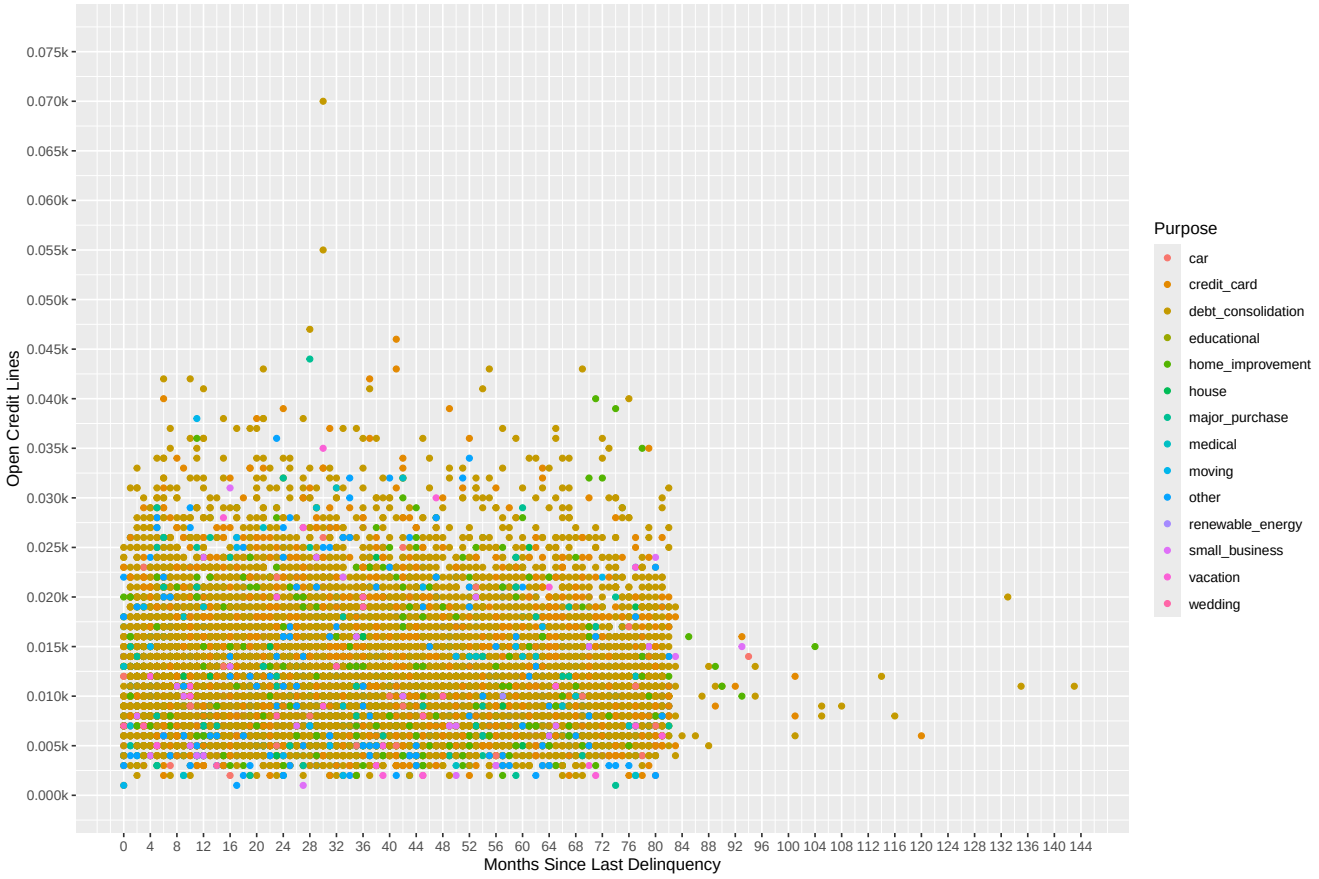

Figure 14: Scatterplot shwoing the relationship between the months since last delinquency and the open credit lines colored by purpose.

```r
# Create scatterplot with mean and median lines, correlation coefficient, and trendline
scatterplot_mths_since_last_delinq_open_acc_emp_length <- ggplot(myLCdata, aes(x = mths_since_last_delin
  geom_point() +
  scale_x_continuous(breaks = seq(0, 144, by = 4), limits = c(0, 144)) +
  scale_y_continuous(breaks = seq(0, 80, by = 5), labels = scales::unit_format(unit = "k", scale = 1e-3,
  labs(x = "Months Since Last Delinquency", y = "Open Credit Lines", color = "Employment Length")

scatterplot_mths_since_last_delinq_open_acc_emp_length
```

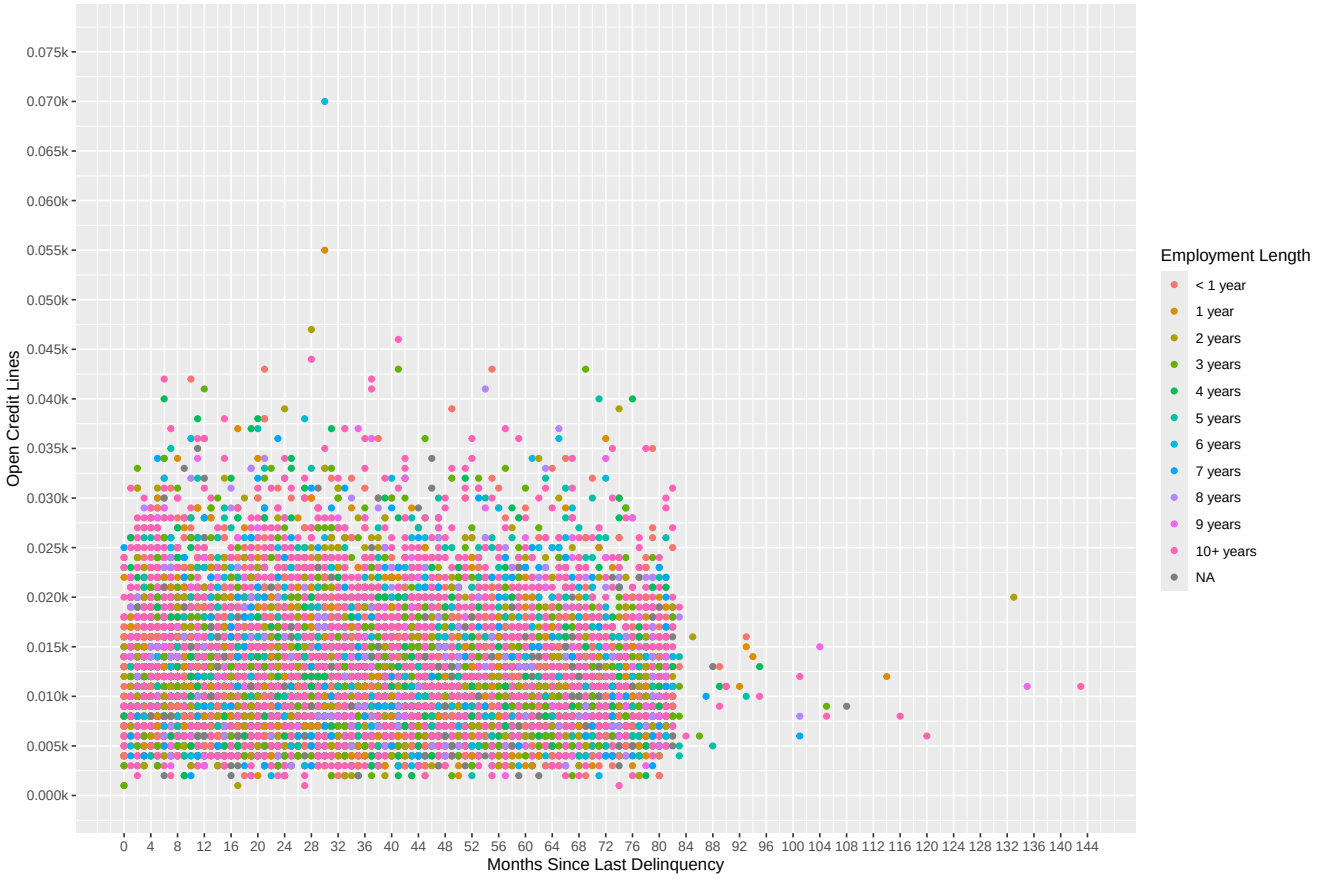

Figure 15: Scatterplot shwoing the relationship between the months since last delinquency and the open credit lines colored by employment length.

To assess whether adding a third attribute via the color channel would provide new insights, a scatterplot of 'months since last delinquency' and 'open credit lines' was created. Three attributes were added via color: 'total current balance' in Figure 13, 'purpose' in Figure 14, and 'employment length' in Figure 15. Unfortunately, none of these plots revealed any new insights or knowledge, as the added attributes visualized by color did not uncover discernible clusters, patterns, or trends.

Months Since Last Delinquency & Total Current Balance

```
# Create scatterplot with mean and median lines, correlation coefficient, and trendline
scatterplot_mths_since_last_delinq_tot_cur_bal_emp_length <- ggplot(myLCdata, aes(x = mths_since_last_de
  geom_point() +
  scale_x_continuous(breaks = seq(0, 144, by = 4), limits = c(0, 144)) +
  scale_y_continuous(breaks = seq(0, 4200000, by = 100000), labels = scales::unit_format(unit = "k", sca
  labs(x = "Total Current Balance in $k", y = "Months Since Last Delinquency", color = "Employment Lengt

scatterplot_mths_since_last_delinq_tot_cur_bal_emp_length
```

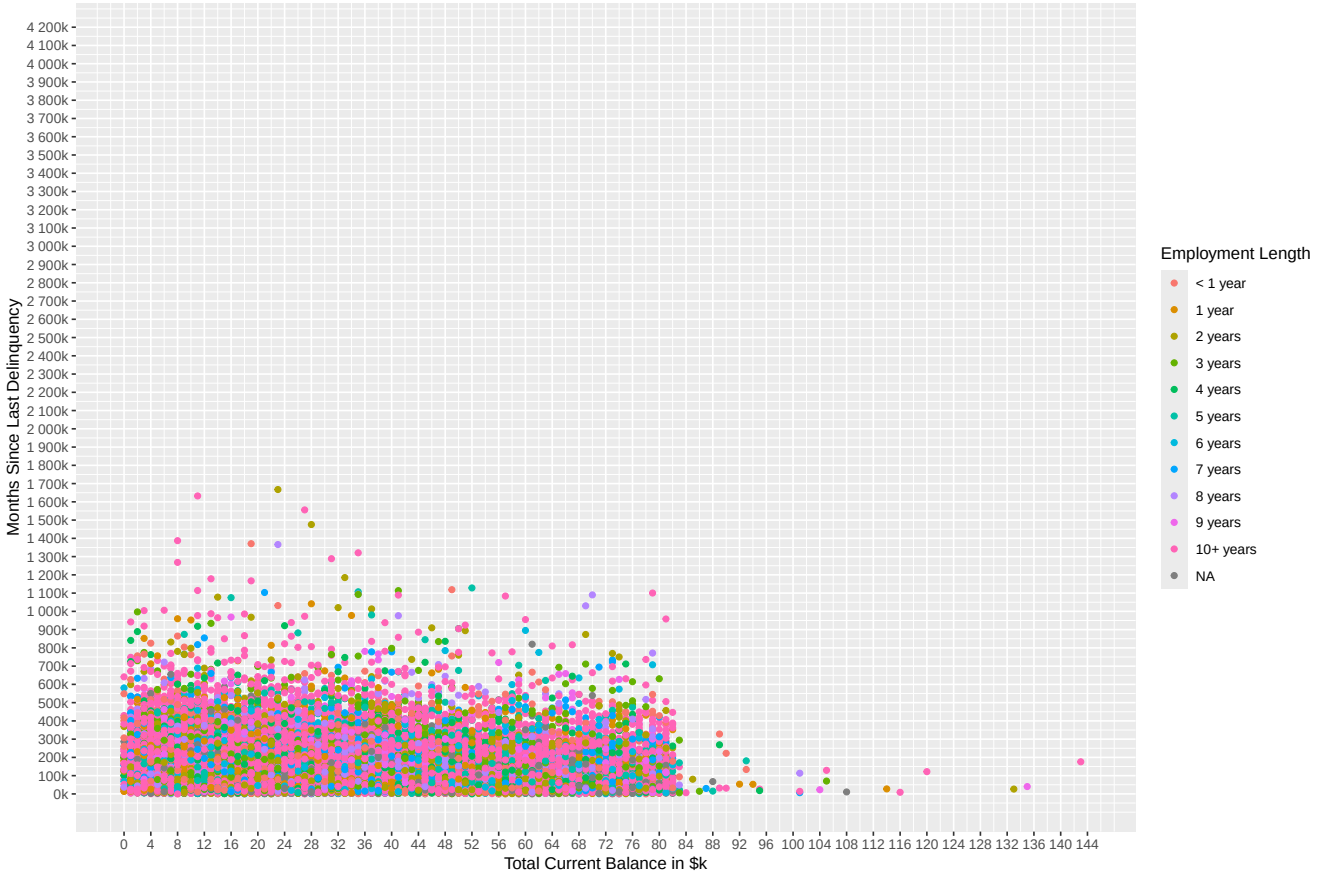

Figure 16: Scatterplot shwoing the relationship between the months since last delinquency and the total current balance colored by employment length.

```r
# Define the breaks for the color scale
color_breaks <- seq(0, 80, by = 10)  # Change the step size to 10

# Create scatterplot with mean and median lines, correlation coefficient, and trendline
scatterplot_mths_since_last_delinq_tot_cur_bal_open_acc <- ggplot(myLCdata, aes(x = mths_since_last_deli
  geom_point() +
  scale_x_continuous(breaks = seq(0, 144, by = 4), limits = c(0, 144)) +
  scale_y_continuous(breaks = seq(0, 4200000, by = 100000), labels = scales::unit_format(unit = "k", sca
  scale_color_continuous(breaks = color_breaks, low = "#4B0082", high = "#DBB7FF") +  # Specify the low
  labs(x = "Total Current Balance in $k", y = "Months Since Last Delinquency", color = "Open Credit Line

scatterplot_mths_since_last_delinq_tot_cur_bal_open_acc
```

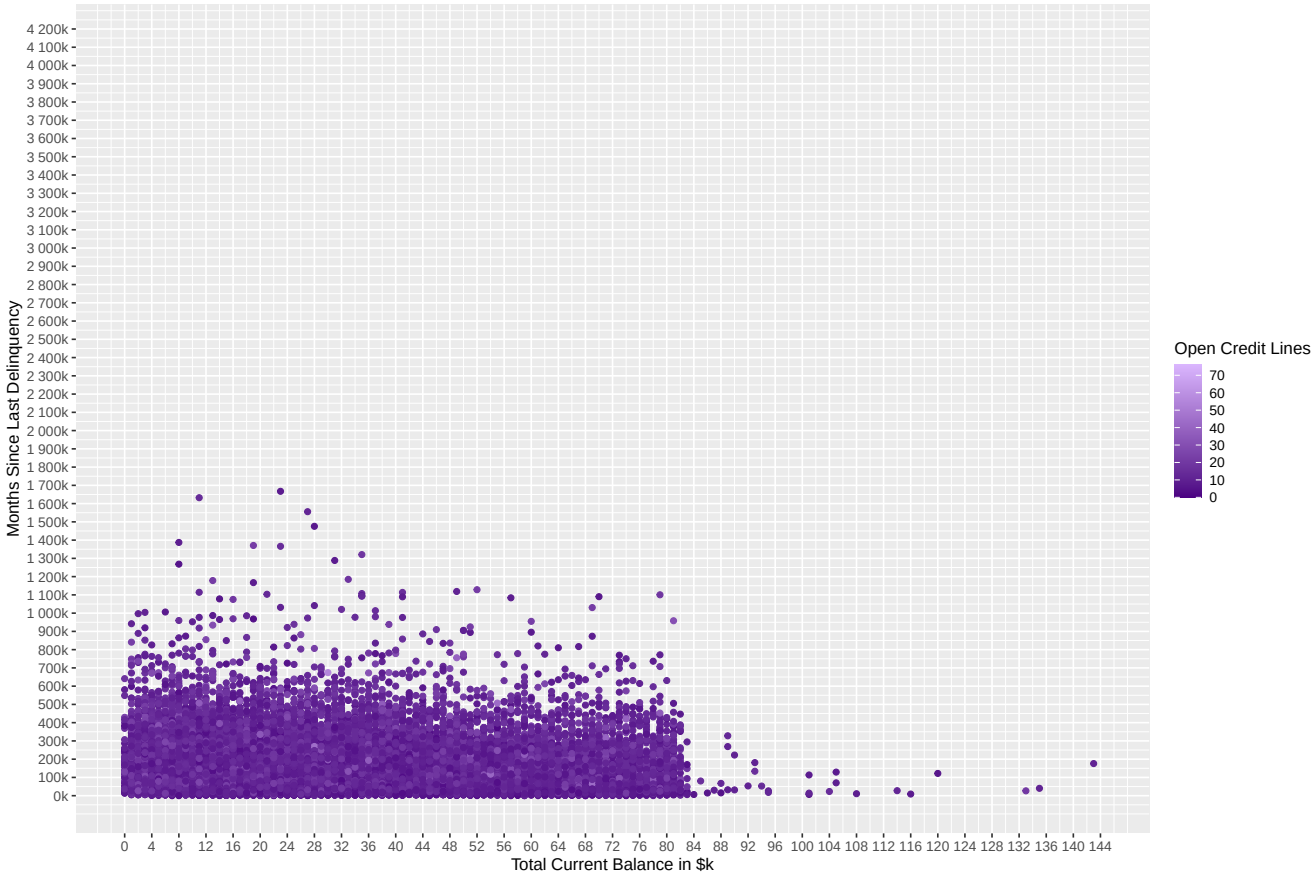

Figure 17: Scatterplot shwoing the relationship between the months since last delinquency and the total current balance colored by open credit lines.

```r
# Create scatterplot with mean and median lines, correlation coefficient, and trendline
scatterplot_mths_since_last_delinq_tot_cur_bal_purpose <- ggplot(myLCdata, aes(x = mths_since_last_delin
  geom_point() +
  scale_x_continuous(breaks = seq(0, 144, by = 4), limits = c(0, 144)) +
  scale_y_continuous(breaks = seq(0, 4200000, by = 100000), labels = scales::unit_format(unit = "k", sca
  labs(x = "Total Current Balance in $k", y = "Months Since Last Delinquency", color = "Purpose")

scatterplot_mths_since_last_delinq_tot_cur_bal_purpose
```

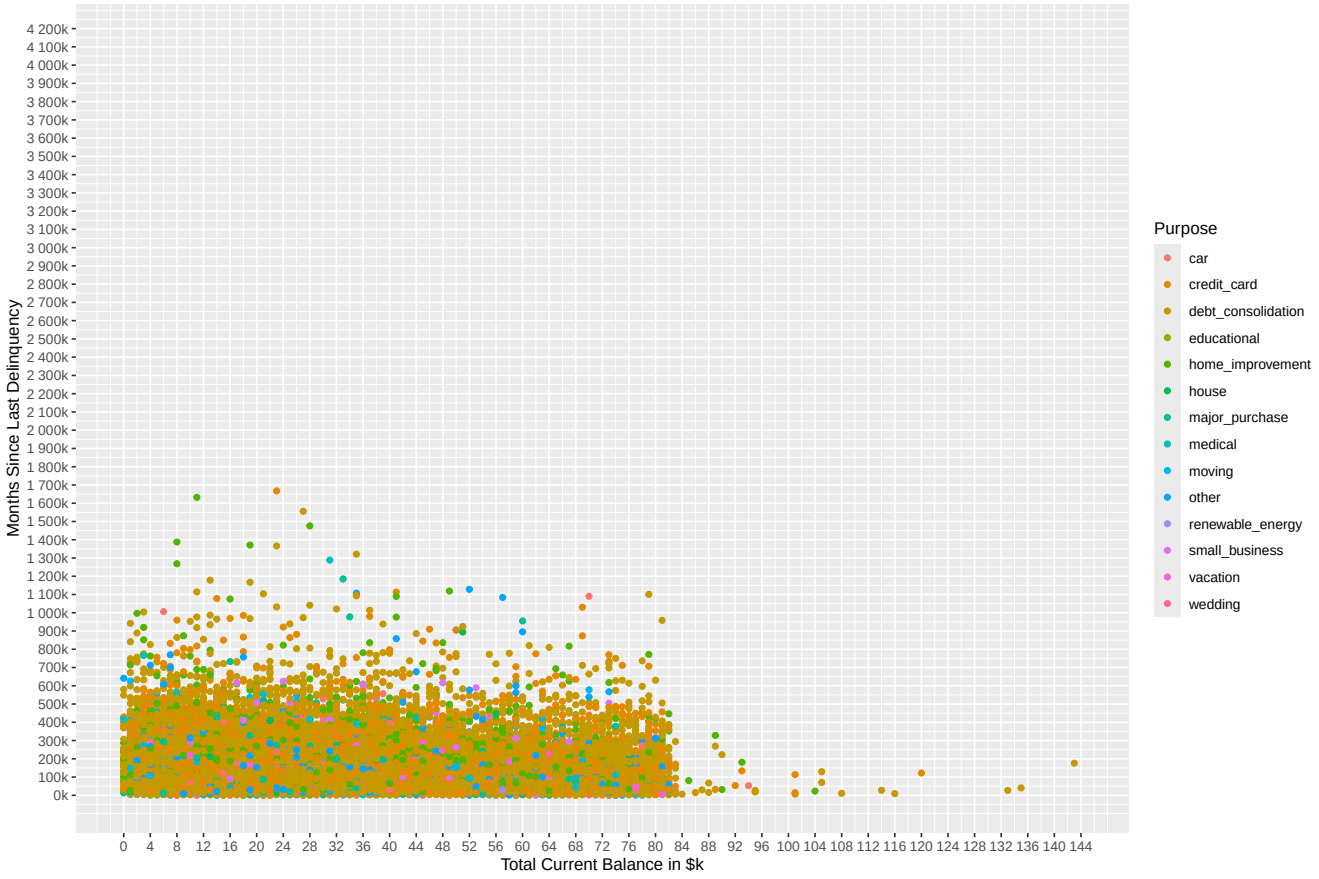

Figure 18: Scatterplot shwoing the relationship between the months since last delinquency and the total current balance colored by purpose.

To further assess whether adding a third attribute via the color channel would provide new insights, a scatterplot of 'months since last delinquency' and 'total current balance' was created. Three attributes were added via color: 'employment length' in Figure 16, 'open credit lines' in Figure 17, and 'purpose' in Figure 18. Unfortunately, none of these plots revealed any new insights or knowledge, as the added attributes visualized by color did not uncover discernible clusters, patterns, or trends.

Interactive 2-dimensional Subplot of Pairsplot with a 3rd Attribute

Due to the PDF format of this quarto file and the configured settings, interactive plots using the `ggplotly()` function from the plotly library are not feasible. The format of this quarto file is optimized for PDF output, and converting it to HTML or other formats that support interactivity would require significant adjustments to its settings and formatting. As a result, interactive exploration of plots is not available in this document.

However, if you wish to view an interactive plot, you can copy the code from one of the following figures into the below code, insert the variable name of the corresponding plot into the code below (uncomment the code first), and then run the code in an R script or quarto file with HTML output. Please also ensure that the data set `myLCdata` is loaded in the environment or copy the code from the preliminaries section into the corresponding file where you want to run the code.

Figure 10 —variable name —> `scatterplot_open_acc_tot_cur_bal_purpose`

Figure 11 —variable name —> `scatterplot_open_acc_tot_cur_bal_emp_length`

Figure 12 —variable name —> `scatterplot_open_acc_tot_cur_bal_mths_since_last_delinq`

Figure 13 —variable name —> `scatterplot_mths_since_last_delinq_open_acc_tot_cur_bal`

Figure 14 —variable name —> `scatterplot_mths_since_last_delinq_open_acc_purpose`

Figure 15 —variable name —> `scatterplot_mths_since_last_delinq_open_acc_emp_length`

Figure 16 —variable name —> `scatterplot_mths_since_last_delinq_tot_cur_bal_emp_length`

Figure 17 —variable name —> `scatterplot_mths_since_last_delinq_tot_cur_bal_open_acc`

Figure 18 —variable name —> `scatterplot_mths_since_last_delinq_tot_cur_bal_purpose`

```
## to uncomment this code highlight it an press Ctrl + Shift + C.
#
## make sure myLCdata is loaded !!!
#
## ======================================
## <code of the desired figure/plot HERE
## or make sure it is loaded in environment>
## ======================================
#
# # Install necessary package
# #install.packages("plotly") # already installed
#
# # Load necessary library
# library(plotly)
# # Convert ggplot to plotly
# interactive_plot <- ggplotly(<HERE variable name of plot>)
#
# # Display the interactive plot
# interactive_plot
```

Further 2-dimensional Plots

To get a deeper insight into the data and explore potential correlations, additional plots have been generated.

Employment Length & Total Current Balance

```
# Convert emp_length to numeric format
emp_length_numeric <- ifelse(grepl("< 1 year", myLCdata$emp_length), 0,
                             ifelse(grepl("10\\+ years", myLCdata$emp_length), 10,
                                    as.numeric(gsub("[^0-9]", "", myLCdata$emp_length))))

# Calculate the correlation coefficient
correlation_emp_tot_cur_bal <- cor(emp_length_numeric, myLCdata$tot_cur_bal, use = "complete.obs")

# Print the correlation coefficient
print(correlation_emp_tot_cur_bal)
```

```
[1] 0.09871345
```

```r
# Sort employment length categories
myLCdata$emp_length <- factor(myLCdata$emp_length, levels = c("< 1 year", "1 year", "2 years", "3 years"

# Create the scatter plot with trend line
scatterplot1 <- ggplot(myLCdata, aes(x = emp_length, y = tot_cur_bal)) +
  geom_point(alpha = 0.6) +  # Add points with transparency
  scale_y_continuous(breaks = seq(0, 4200000, by = 100000), labels = scales::unit_format(unit = "k", sca
  labs(title = "Relationship between Employment Length and Total Current Balance",
       x = "Employment Length",
       y = "Total Current Balance in $k") +  # Set axis labels
  theme_minimal()  # Set plot theme

# Display the scatter plot
scatterplot1
```

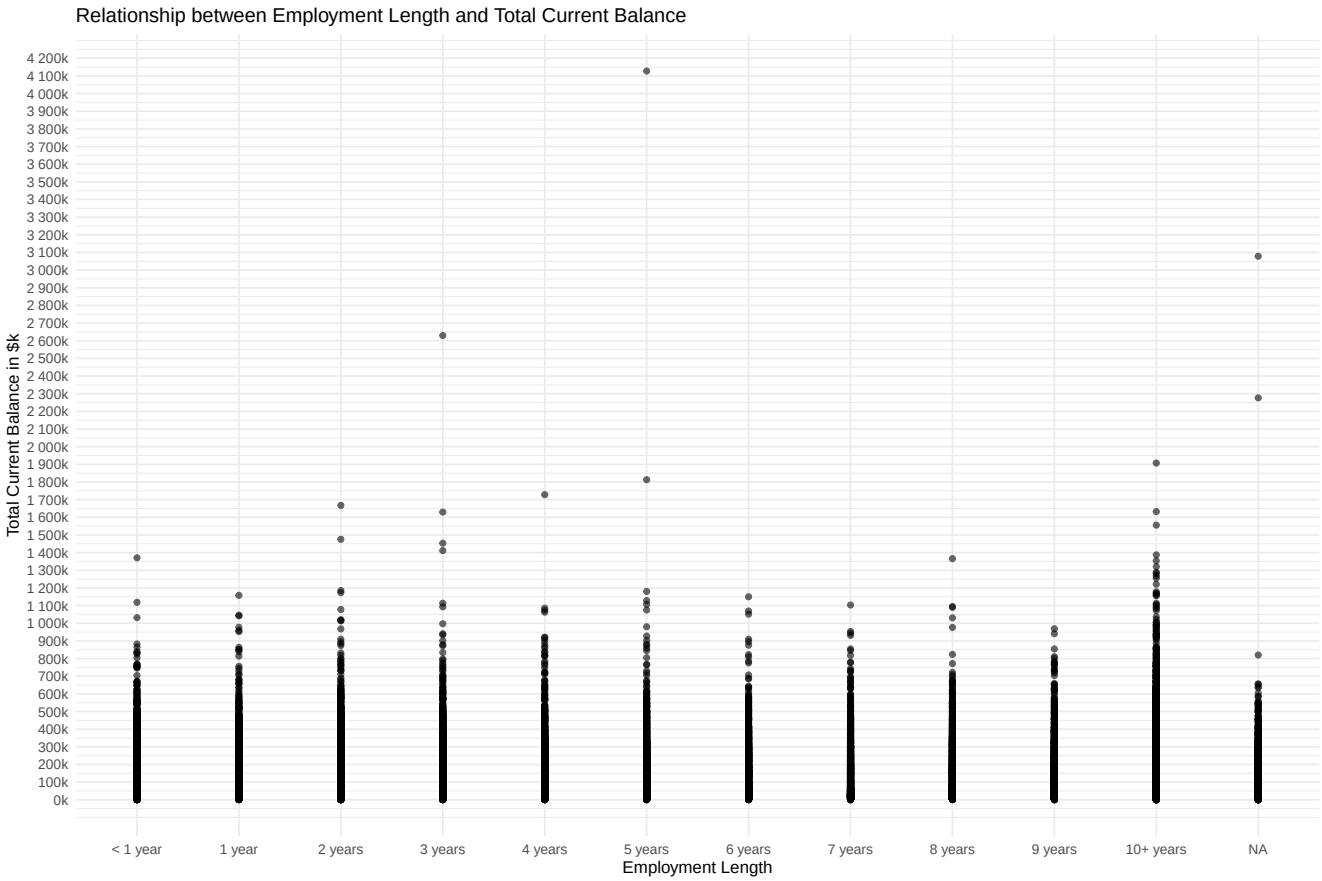

Figure 19: Scatterplot of employment length and total current balance

The correlation between the current total balance and the employment length (after conversion into numerical values) is only 0.0987, which indicates a weak correlation. Furthermore, when looking at Figure 19, it can be seen that the scatter plot looks more like a histogram with bars. This resemblance is due to the fact that the employment length attribute is of the character type and therefore consists of categories. It is noticeable that the category "10+ years" has the highest number of loan records. However, no significant or novel insights can be gained from these observations.

Purpose & Months Since Last Delinquency

```
# Calculate count of each purpose category
purpose_counts <- myLCdata %>%
  group_by(purpose) %>%
  summarize(count = n()) %>%
  arrange(count)

# Create the box plot with adjusted x-axis breaks and sorted y-axis categories
boxplot1 <- ggplot(myLCdata, aes(x = mths_since_last_delinq, y = factor(purpose, levels = purpose_counts
  geom_boxplot(fill = "skyblue", color = "black", alpha = 0.8) +  # Box plot aesthetics
  labs(title = "Relationship between Months Since Last Delinquency and Purpose of Loan",
       x = "Months Since Last Delinquency",
       y = "Purpose of Loan") +  # Axis labels swapped
  theme_minimal() +  # Set plot theme
  theme(axis.text.y = element_text(angle = 0, hjust = 1)) +
  scale_x_continuous(breaks = seq(0, (max(myLCdata$mths_since_last_delinq, na.rm = TRUE) + 10), by = 4))

# Display the box plot with adjusted x-axis breaks and sorted y-axis categories
boxplot1
```

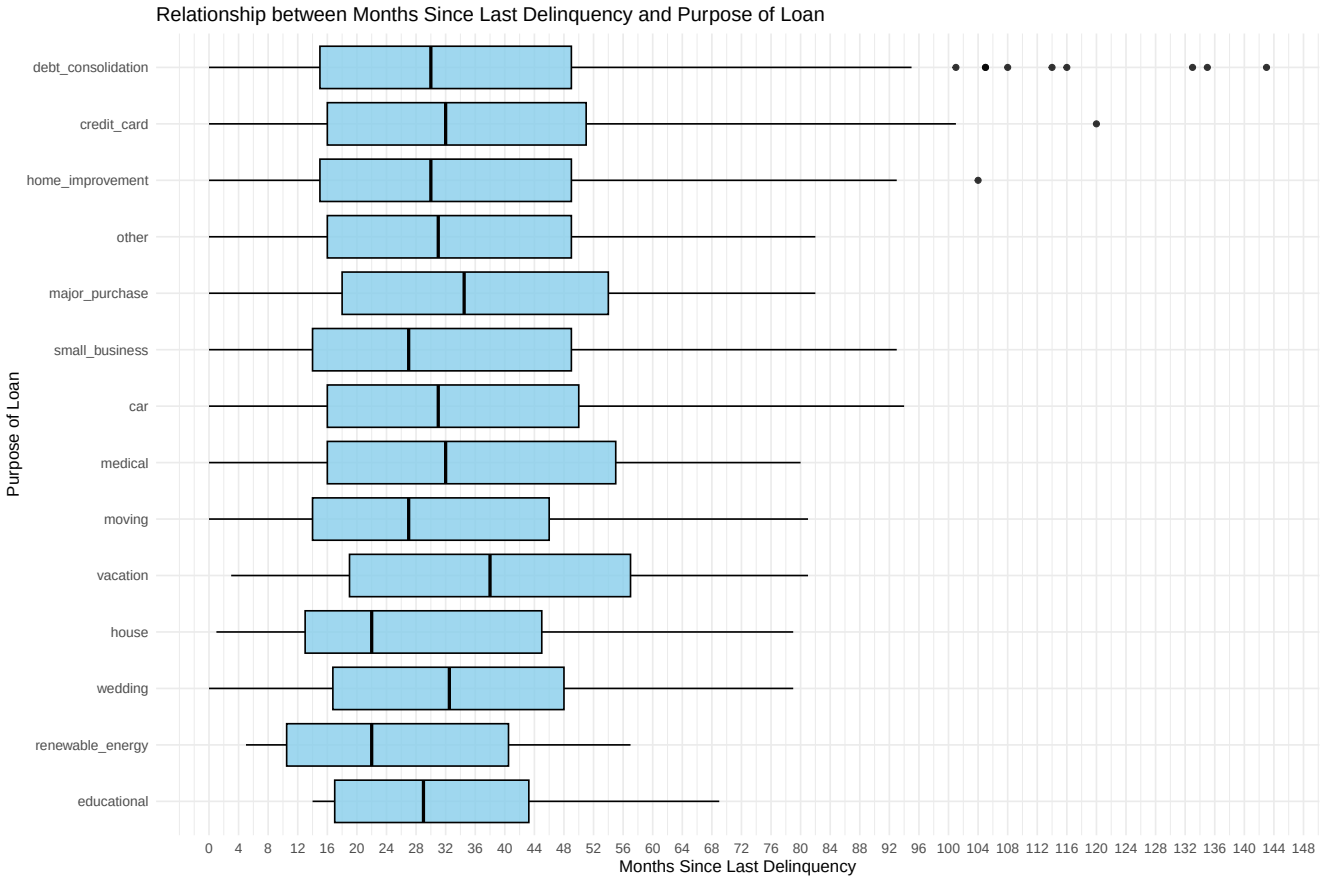

Figure 20: Boxplot of purpose and months since the last delinquency

The boxplots in Figure 20 seems quite symmetric, what indicates a symmetric distribution and therefore a balanced data spread. The categories of purpose are sorted according to the count of loan records increasingly per category in the plot. A potential correlation could be interpreted, suggesting that as loan records increase per purpose category, the range of months since last delinquency also tends to increase, with higher numbers of months included. Notably, the categories "debt consolidation" and "credit card" show the longest boxplots and the highest count of loan records, which indicates a longer and higher range of months since the last delinquency. Outliers can be observed in these categories, while other categories seem to have no outliers. On the other hand, the "educational" and "renewable energy" categories have the lowest count of loan records and shortest boxplots, indicating a shorter and smaller range of months since the last delinquency. Overall, no significant or novel insights can be gained from these observations.

Purpose & Total Current Balance

```
# Calculate count of each purpose category
purpose_counts <- myLCdata %>%
  group_by(purpose) %>%
  summarize(count = n()) %>%
  arrange(count)

# Create the box plot
boxplot2 <- ggplot(myLCdata, aes(x = factor(purpose, levels = purpose_counts$purpose), y = tot_cur_bal))
  geom_boxplot(fill = "skyblue", color = "black", alpha = 0.8) +  # Box plot aesthetics
  scale_y_continuous(labels = scales::unit_format(unit = "k", scale = 1e-3, sep = "")) +    # Format y-a
  labs(title = "Relationship between Purpose of Loan and Total Current Balance",
       x = "Purpose of Loan",
       y = "Total Current Balance in $k") +  # Axis labels
  theme_minimal() +  # Set plot theme
  theme(axis.text.x = element_text(angle = 90, hjust = 1))  # Rotate x-axis labels for better readabilit

# Display the box plot
boxplot2
```

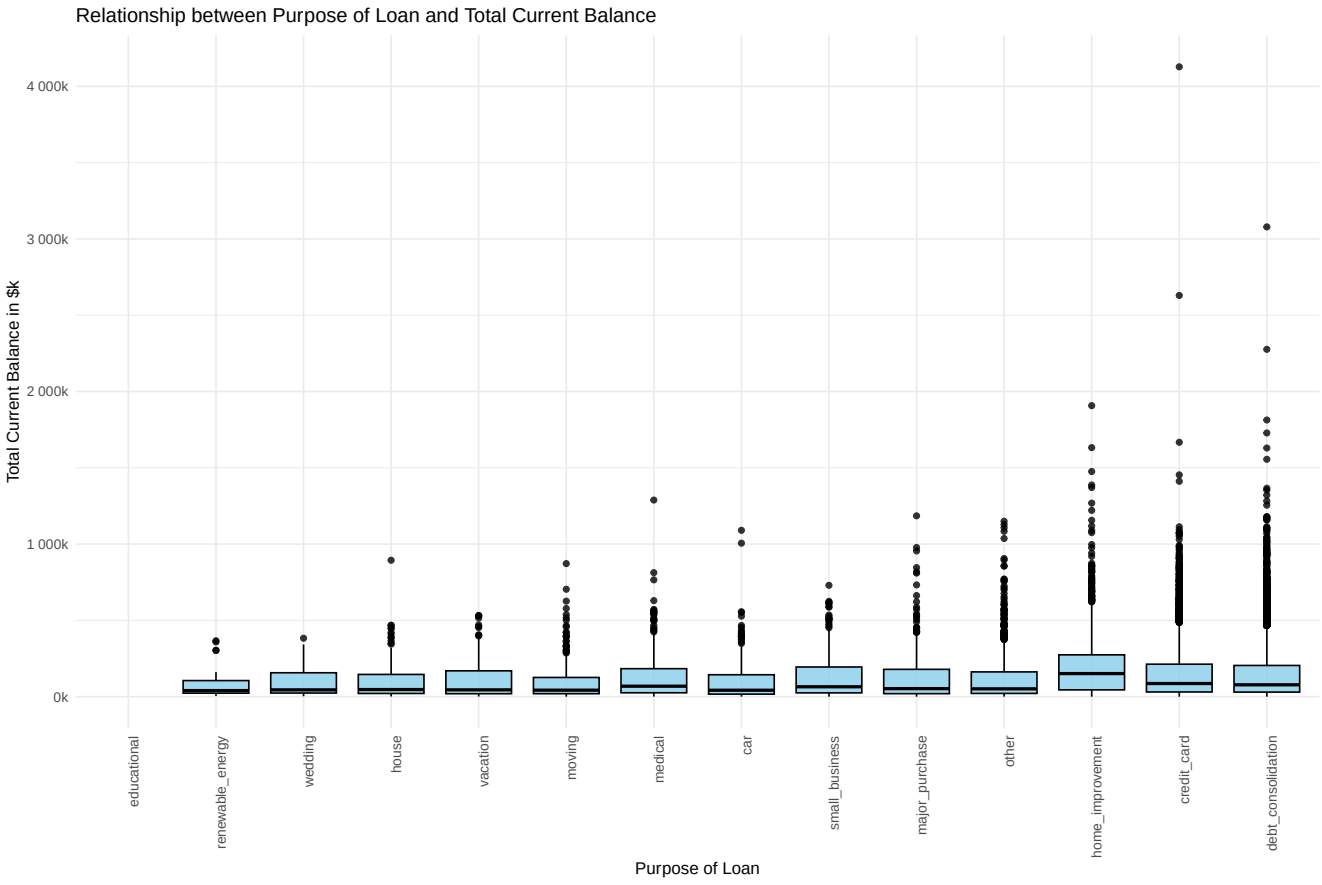

Figure 21: Boxplot of purpose and total current balance

```r
# Calculate count of each purpose category
purpose_counts <- myLCdata %>%
  group_by(purpose) %>%
  summarize(count = n()) %>%
  arrange(count)

# Create scatter plot with trend line
scatterplot2 <- ggplot(myLCdata, aes(x = tot_cur_bal, y = factor(purpose, levels = purpose_counts$purpos
  geom_point() +
  scale_x_continuous(labels = scales::unit_format(unit = "k", scale = 1e-3, sep = "")) +    # Format x-a
  labs(title = "Relationship between Total Current Balance and Purpose of Loan",
       x = "Total Current Balance in $k",
       y = "Purpose of Loan") +
  theme_minimal()

# Display scatter plot
scatterplot2
```

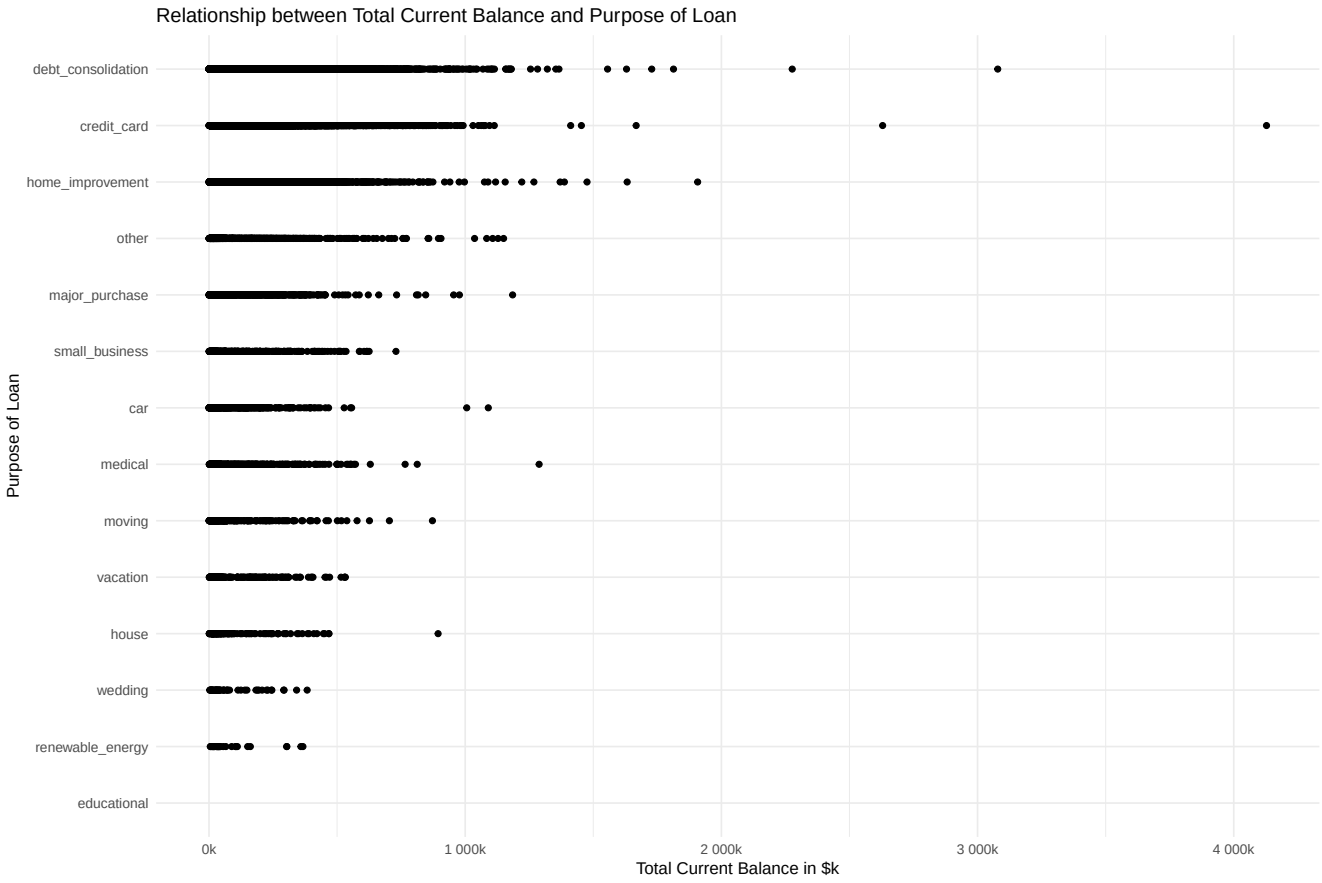

Figure 22: Scatterplot of purpose and total current balance

When looking at Figure 21 and Figure 22, where the purpose categories are sorted based on the count of loan records per category, a possible trend occurs: as the count of loan records increases, so does the total current balance (Figure 22) and, outliers tend to occur at higher total current balances (Figure 21). Otherwise, no significant or novel insights can be gained from these observations.

Employment Length & Open Credit Lines

```
# Convert emp_length to numeric format
emp_length_numeric <- ifelse(grepl("< 1 year", myLCdata$emp_length), 0,
                             ifelse(grepl("10\\+ years", myLCdata$emp_length), 10,
                                    as.numeric(gsub("[^0-9]", "", myLCdata$emp_length))))

# Calculate the correlation coefficient
correlation_emp_open_acc <- cor(emp_length_numeric, myLCdata$open_acc, use = "complete.obs")

# Print the correlation coefficient
print(correlation_emp_open_acc)
```

```
[1] 0.04473438
```

```r
# Create the scatter plot
scatterplot3 <- ggplot(myLCdata, aes(x = emp_length, y = open_acc)) +
  geom_point(color = "skyblue") +  # Scatter plot aesthetics
  labs(title = "Relationship between Employment Length and Open Credit Lines",
       x = "Employment Length",
       y = "Open Credit Lines") +  # Axis labels
  theme_minimal() +  # Set plot theme
  theme(axis.text.x = element_text(angle = 45, hjust = 1))  # Rotate x-axis labels for better readabilit

# Display the scatter plot
scatterplot3
```

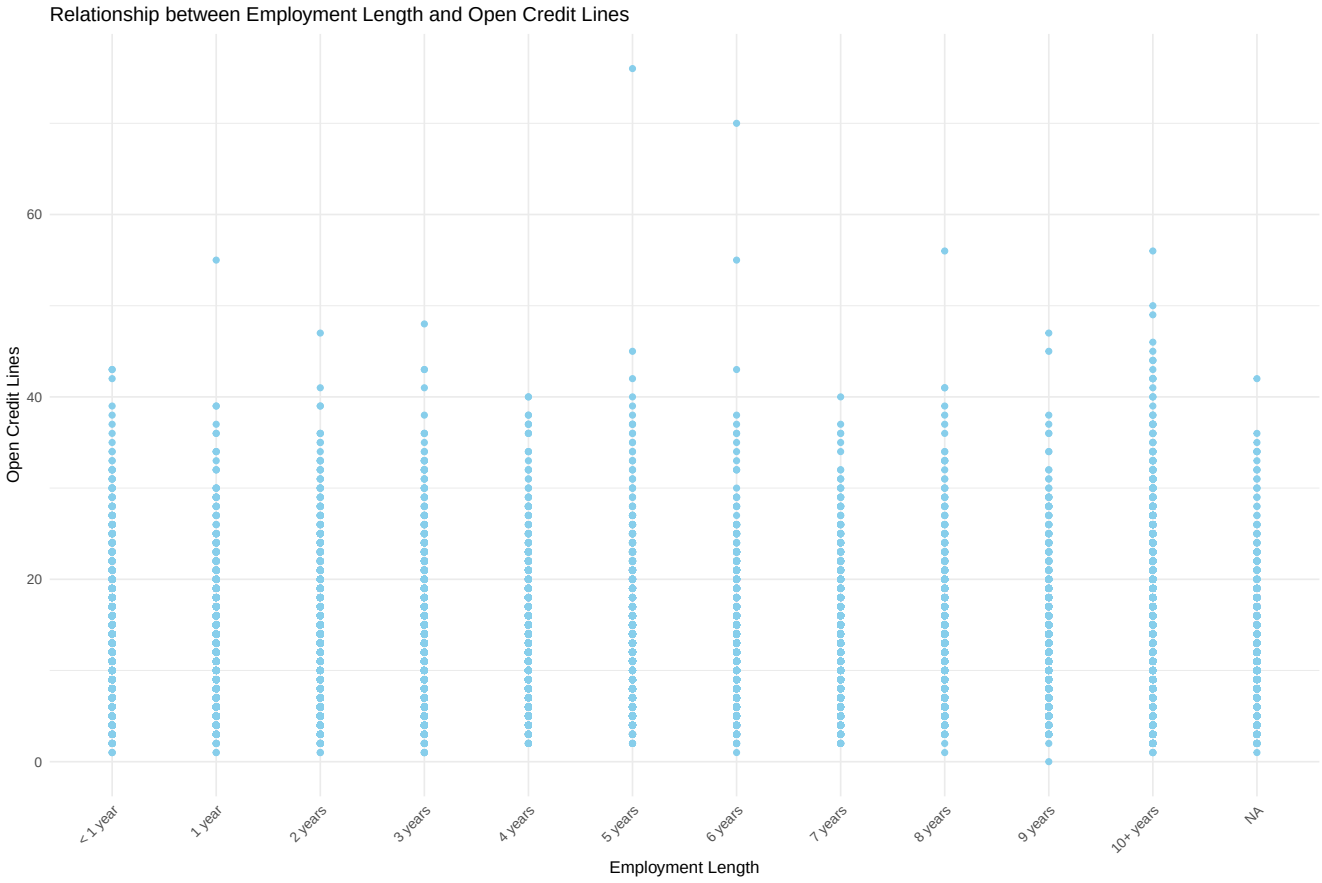

Figure 23: Scatterplot of employment length and open credit lines

The correlation between the employment length (after conversion into numerical values) and the open credit lines is only 0.0447, which indicates a weak correlation. Furthermore, when looking at Figure 23, it can be seen that the scatter plot looks more like a histogram with bars. This resemblance is due to the fact that the employment length attribute is of the character type and therefore consists of categories. No significant or novel insights can be gained from these observations.