2.3 Muhammad Noorullah Baig (TP077979)

For the purpose of this analysis, the cleaned dataset prepared by the group which is named credit_risk, was assigned to data. This ensures consistency in variable naming and ease of handling throughout the modelling and visualization process.

```
data = credit_risk
```

Figure 2.3.0

Objective 1: To investigate the impact of *personal_status* on credit risk classification.

Financial institutions assess a variety of personal and professional factors when determining an individual's credit classification. They often bank on analyzing trends and patterns associated with different life circumstances and their effects on an individual's ability to pay back the loan based on previous data. Our focus in the current analysis will be to find out whether an individual's gender and marital status serve as significant indicators that can be used to classify credit risk.

Analysis Questions:

Analysis Question 1: How is the dataset distributed for the ***personal_status*** column? What are some noticeable trends?

A good first step when trying to understand a dataset is always a holistic summary of the values. A table of all the unique values in the ***personal_status*** column and their frequencies are given below:

```
> data$personal_status = as.factor(data$personal_status)
> summary(data$personal_status)
female div/dep/mar       male div/sep       male mar/wid        male single
              1932                517                768               2783
```

Figure 2.3.1

female div/dep/mar	male div/sep	male mar/wid	male single
1932	517	768	2783

Table 2.3.1

We can better visualize the distribution of these values using a pie chart which is given below:

```
#Pie chart showing distribution of personal status
status_counts <- data %>%
  group_by(personal_status) %>%
  summarise(Frequency = n(), .groups = "drop") %>%  # Summarize to get counts for each status
  mutate(Percentage = round(Frequency / sum(Frequency) * 100, 1))  # Add percentage column
status_counts <- status_counts %>%
  mutate(Label = paste(Percentage, "%"))
ggplot(status_counts, aes(x = "", y = Frequency, fill = personal_status)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y") +
  labs(title = "Distribution of Personal Status",
       x = NULL, y = NULL, fill = "Personal Status") +
  geom_text(aes(label = Label), position = position_stack(vjust = 0.5), size = 4) +
  theme_void() +
  theme(legend.title = element_text(size = 12),
        legend.text = element_text(size = 10),
        plot.title = element_text(size = 14, face = "bold", hjust = 0.5))
```

Figure 2.3.2

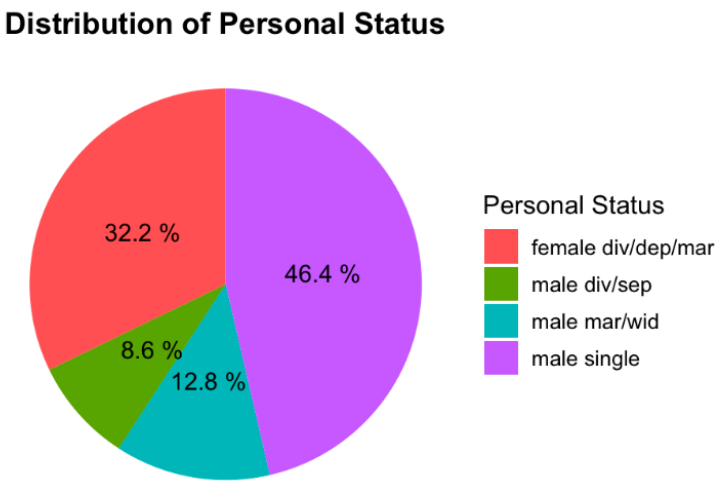

Figure 2.3.3

The pie chart illustrates the disproportionate representation of single males in the dataset, representing 46.4% of the total. Married/divorced/separated females collectively account for 32.2%, followed by married/widowed males at 12.8%. While divorced/separated males make up the remaining portion. This dataset does not have any representation of single females which can be problematic for our analysis.

To further explore the relationship between these categories with respect to credit risk, a bar chart examining the credit risk classification within each group is created.

```r
#Preparing data for visualization
summary_data <- data %>%
  group_by(personal_status, class) %>%
  summarise(Frequency = n(), .groups = "drop") %>%
  mutate(Proportion = Frequency / sum(Frequency),
         Percentage = round(Frequency / sum(Frequency) * 100, 1),
         Label = paste0(Percentage, "%"))

#Frequency bar chart for personal status
ggplot(summary_data, aes(x = personal_status, y = Frequency, fill = class)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Frequency of Good and Bad Credit Ratings by Personal Status",
       x = "Personal Status",
       y = "Frequency",
       fill = "Credit Class") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Figure 2.3.4

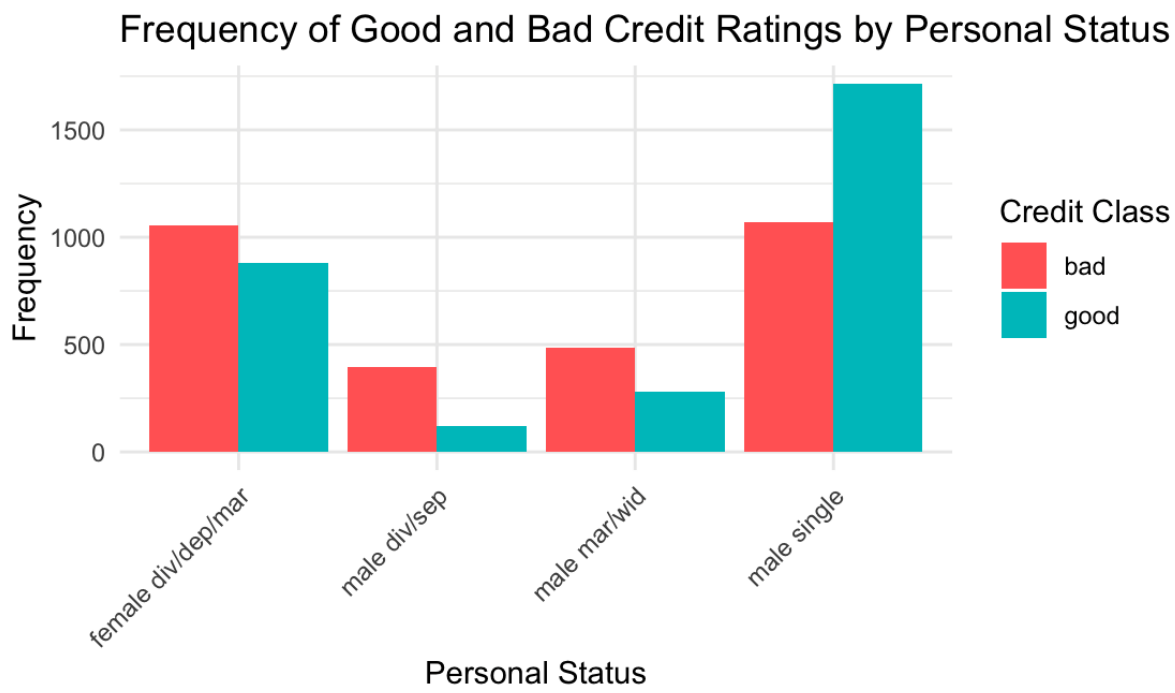

Figure 2.3.5

The frequency bar chart again gives us a good overview of the distribution of credit classes among the categories. One interesting contrast to note is that single males have the highest frequency of good credit risk while the divorced and separated male category has the least frequency of good credit risk.

We will make a proportion chart for the frequency to emphasize relative impacts of the categories on credit risk classification.

```r
#Proportion Bar Chart of personal status including credit class
ggplot(summary_data, aes(x = personal_status, y = Proportion, fill = class)) +
  geom_bar(stat = "identity", position = "fill") +
  geom_text(aes(label = scales::percent(Proportion)),
            position = position_fill(vjust = 0.5), size = 3) +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Proportion of Credit Ratings by Personal Status",
       x = "Personal Status",
       y = "Proportion",
       fill = "Credit Class") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Figure 2.3.6

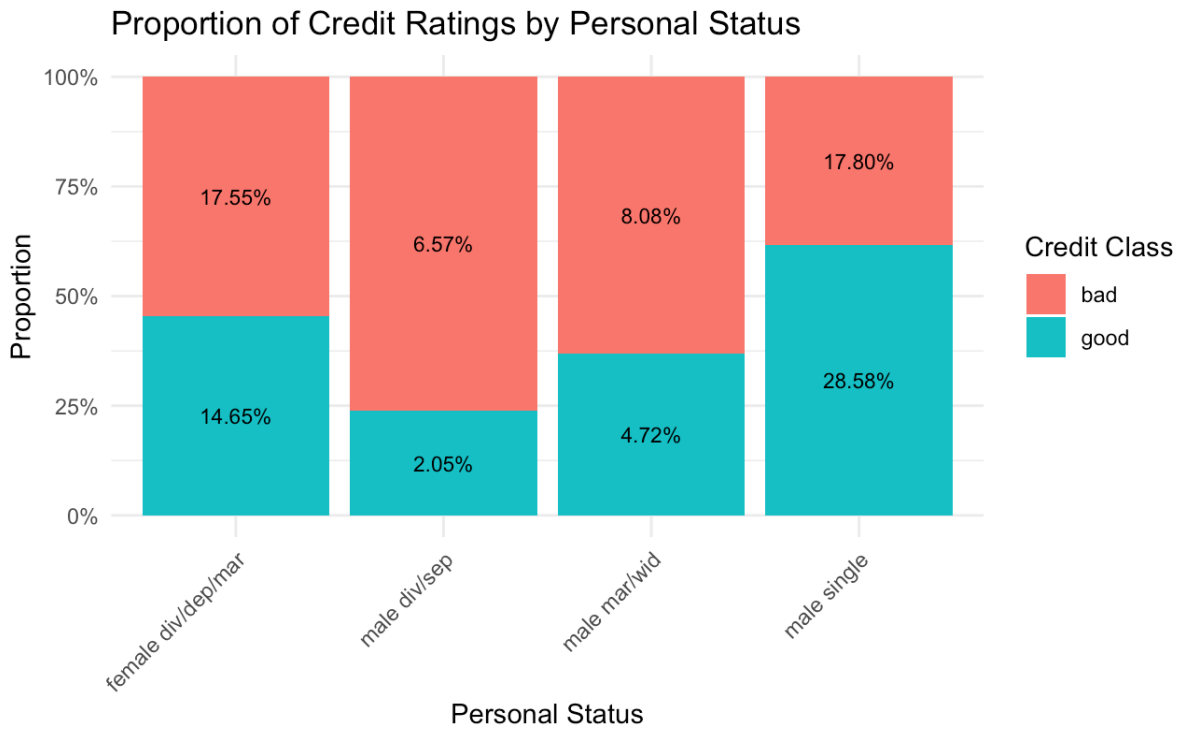

Figure 2.3.7

The above stacked bar chart further emphasizes our observation of single males being the most dominant in having good credit risk while divorced and separated males are the most dominant in having the least proportion of credit risk.

Analysis Question 2: Which demographic factor, gender or marital status, should be further pursued as a predictor for credit risk classification?

There are two ways to approach this data. One is to analyze the effect of marital status on credit risk classification, and the other is to focus on gender. However, as noted previously,

our dataset lacks representation for single females. Therefore, it makes more sense to pursue the latter option.

We will now proceed to derive a new ***gender*** column from ***personal_status*** to investigate its impact on credit risk classification.

```
> #Creating a new column called "Gender"
> data$gender <- factor(ifelse(grepl("^female", credit_risk$personal_status), "female", "male"))
> summary(data$gender)
female   male
  1932   4068
```

Figure 2.3.8

Analysis Question 3: How does the distribution of credit risk classifications vary between the genders, and are any disparities between credit risk classification present between the genders?

The distribution of values in the new gender column are given in the table below:

Female	Male
1932	4068

Table 2.3.2

We now form a 3d pie chart using the "plotrix" library to visualize the above-mentioned distribution of the "gender" column.

```
# Pie chart to show gender distribution
gender_counts <- table(data$gender)
percentages <- round(prop.table(gender_counts) * 100, 1)
library(plotrix)
pie3D(gender_counts,
      labels = paste(percentages, "%", "(", as.numeric(gender_counts), ")"),
      explode = 0.2,
      col = c("pink", "skyblue"),
      radius = 1.0,
      labelcex = 1.0,
      theta = 0.6)
title(main = "Distribution of Gender", cex.main = 1.5)
legend("topright",
       legend = names(gender_counts),
       fill = c("pink", "skyblue"),
       title = "Gender",
       cex = 1)
```

Figure 2.3.9

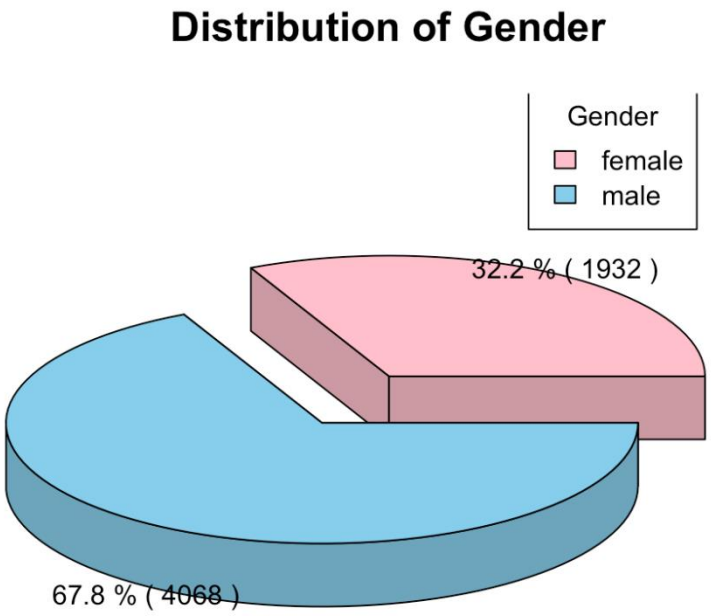

Figure 2.3.10

The 3d pie chart shows as expected demonstrates the disproportionate number of males present in the dataset in comparison to females. It clearly shows that males form the bulk of the sample being 67.8% while females form the rest of the 32.2% of the sample. This disproportion is bad for our analysis as this introduces bias and generalization in our analysis.

Analysis Question 4: What are the proportions of individuals with bad and good credit risk based on gender?

We can inspect the proportions of individuals having good and bad credit risk according to their gender using a stacked bar chart.

```r
#Preparing data for chart
count_gender <- data %>%
  count(gender, class) %>%
  group_by(gender) %>%
  mutate(Proportion = n / sum(n))

#Proportion Bar Chart with gender including credit class
ggplot(count_gender, aes(x = gender, y = Proportion, fill = class)) +
  geom_bar(stat = "identity", position = "fill") +
  geom_text(aes(label = scales::percent(Proportion)),
            position = position_fill(vjust = 0.5), size = 3) +
  scale_y_continuous(labels = scales::percent) +
  labs(title = "Proportion of Credit Ratings by Gender",
       x = "Gender",
       y = "Proportion",
       fill = "Credit Class") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Figure 2.3.11

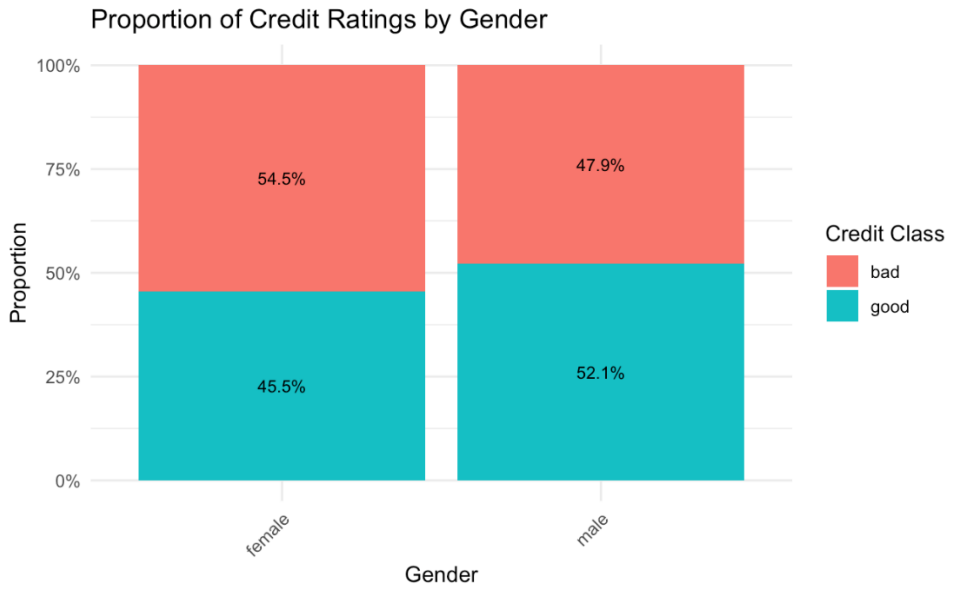

Figure 2.3.12

This stacked bar chart illustrates the proportion of good and bad credit ratings across the two genders. It shows that for both males and females, the proportions of "good" and "bad" credit ratings are visually similar, with a slight difference in the balance between the two classes. While this proportional view provides an overview of the credit classification distribution by

gender, we cannot conclude through this whether these differences are statistically significant.

Literature Review:

A 2013 study on credit card history revealed that women were 26% more likely than men to incur late fees on their credit cards (Mottola, 2013). The study continues to explain that this discrepancy stems from differences in financial behaviors, financial literacy and financial decision-making habits of males and females.

Thus this study gives us room to infer that women are potentially more likely to exhibit behaviors that can cause them to be considered as a bad credit risk.

Hypothesis:

Null Hypothesis (H_0): An individual's gender, specifically being female, does not increase the probability of them being classified as a bad credit risk.

Alternative Hypothesis (H_1): An individual's gender being female increases the probability of them being classified as a bad credit risk.

HYPOTHESIS TESTING:

We begin to test our hypothesis by conducting a chi-square test. A chi-square test is used to measure if two categorical variables are independent of each other.

```
> gendertable = table(data$gender,data$class)
> chisq.test(gendertable)

        Pearson's Chi-squared test with Yates' continuity correction

data:  gendertable
X-squared = 22.848, df = 1, p-value = 1.753e-06
```

Figure 2.3.13

Our chi-square test gives a p-value of 1.753e-06 which is lower than 0.001 making the null hypothesis practically very unlikely. This provides a very strong proof of the relationship between gender and class being significant.

Since our goal is to investigate whether there's a difference in the likelihood of being classified as a bad credit risk between males and females. Therefore, a two-sample test for the

equality of proportions is a suitable choice. This test will help us assess if the observed difference in proportions between the genders is statistically significant or simply due to random chance.

```
> # Perform Two-Proportion Z-Test
> test_result <- prop.test(
+   x = gender_stats$bad_count,
+   n = gender_stats$total_count
+ )
> test_result

        2-sample test for equality of proportions with continuity correction

data:  gender_stats$bad_count out of gender_stats$total_count
X-squared = 22.848, df = 1, p-value = 1.753e-06
alternative hypothesis: two.sided
95 percent confidence interval:
 0.03904139 0.09379359
sample estimates:
   prop 1    prop 2
0.5450311 0.4786136
```

Figure 2.3.14

This p-value extracted is the same as the one from the chi-square test, it confirms that one gender has a higher proportion of bad credit risk than the other. The 95% confidence interval having a range from 0.03904139-0.09379359 suggests that we are 95% confident that the true difference in proportions of bad credit risks between the two genders lies within this interval. Whereas the sample estimates indicate that females are more likely to be classified as bad credit risk (54.5%) compared to males (47.9%).

This test reinforces and quantifies the difference in proportion between genders being classified as bad credit risk.

We proceed to visualize this relationship with a lollipop chart:

```r
#Confidence Interval Lollipop Chart
gender_stats <- data %>%
  group_by(gender) %>%
  summarise(
    bad_count = sum(class == "bad"),
    total_count = n(),
    proportion = bad_count / total_count,
    ci = list(prop.test(bad_count, total_count)$conf.int) # Store CI as a list
  ) %>%
  mutate(
    ci_lower = map_dbl(ci, ~ .[1]), # Extract lower CI
    ci_upper = map_dbl(ci, ~ .[2])  # Extract upper CI
  )
ggplot(gender_stats, aes(x = gender, y = proportion)) +
  geom_point(size = 3, color = "red") +
  geom_errorbar(aes(ymin = ci_lower, ymax = ci_upper), width = 0.2, color = "blue") +
  scale_y_continuous(labels = scales::percent) +
  labs(
    title = "Bad Credit Ratings by Gender with 95% Confidence Intervals",
    subtitle = paste("p-value =", signif(test_result$p.value, digits = 3)),
    x = "Gender",
    y = "Proportion with Bad Credit Rating"
  ) +
  theme_minimal()
```

Figure 2.3.15

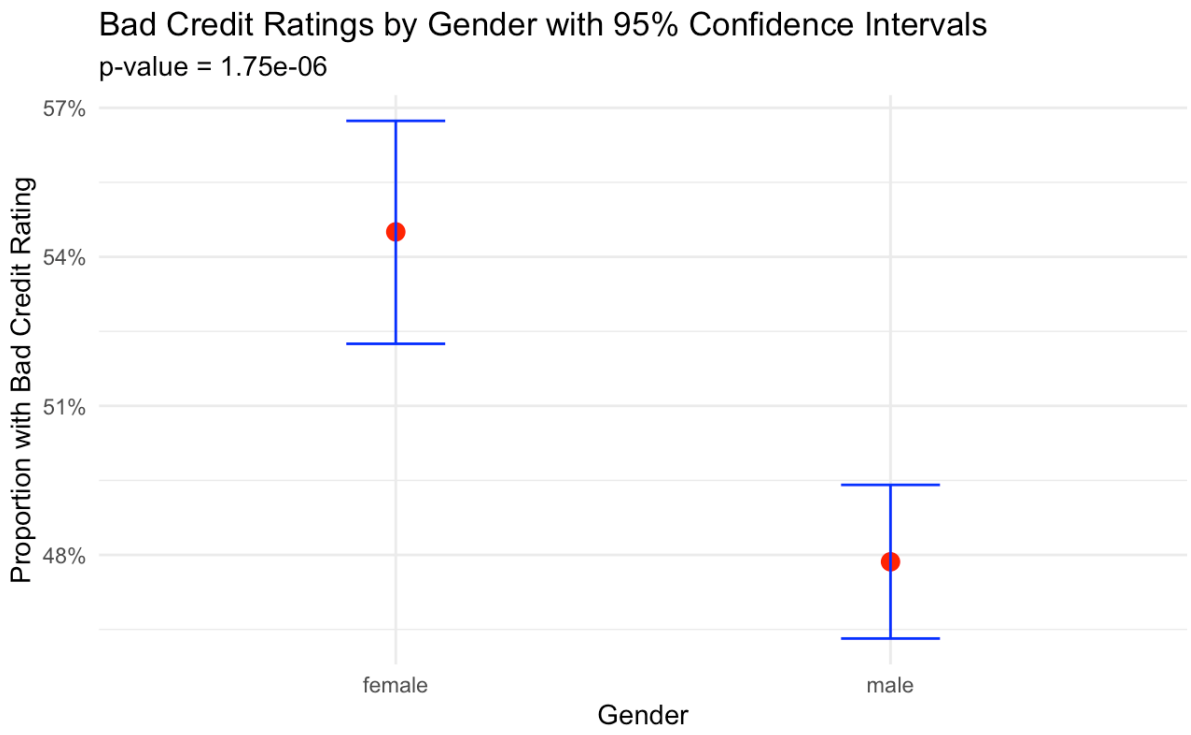

Figure 2.3.16

The lollipop plot above effectively showcases the proportion of bad credit risk for both genders represented by the red dot. The blue bars represent the 95% confidence intervals for each gender's distinct proportion.

From the fact that the two bars do not overlap, we can visually assess that the difference in proportion between the two genders is significant.

Next, to quantify the strength of the relationship, we perform a cramer's v test:

```
> library(vcd)
> assocstats(gendertable)
                     X^2 df   P(> X^2)
Likelihood Ratio 23.137  1 1.5087e-06
Pearson          23.113  1 1.5273e-06

Phi-Coefficient   : 0.062
Contingency Coeff.: 0.062
Cramer's V        : 0.062
```

Figure 2.3.17

The Cramer's V shows an extremely low value of 0.062. While this value does not negate our previous tests, it does contextualize them by showing us that the actual effect of gender on credit classification is weak.

We create a new column derived from the "class" column, consisting of binary values: 0 for bad credit risk and 1 for good credit risk.

```
#Convert class into binary
data <- data %>%
  mutate(class_bin = ifelse(class == "bad", 0, 1))
```

Figure 2.3.18

We create a logistic regression model, which calculates the log-odds of being classified as good credit risk based on gender.

```
#Logistic Regression for gender
logistic_model <- glm(class_bin ~ gender, data = data, family = "binomial")
summary(logistic_model)
```

Figure 2.3.19

```
Call:
glm(formula = class_bin ~ gender, family = "binomial", data = data)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.18061    0.04569  -3.953 7.71e-05 ***
gendermale   0.26621    0.05543   4.803 1.57e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8317.8  on 5999  degrees of freedom
Residual deviance: 8294.6  on 5998  degrees of freedom
AIC: 8298.6

Number of Fisher Scoring iterations: 3
```

Figure 2.3.20

The above logistic regression model shows that the coefficient at the intercept, which represents females, is statistically highly significant having a p-value of 7.71e-05 which is extremely lower than 0.001.

The coefficient at the intercept is negative which means that females have a higher chance of being classified as bad credit risk.

```r
# Add predicted probabilities to the dataset
data$predicted_prob <- predict(logistic_model, type = "response")
prob_summary <- data %>%
  group_by(gender) %>%
  summarise(mean_predicted_prob = mean(predicted_prob),
            .groups = "drop")

#Bar chart to visualize predicted probabilities
ggplot(prob_summary, aes(x = gender, y = mean_predicted_prob, fill = gender)) +
  geom_bar(stat = "identity") +
  scale_y_continuous(labels = scales::percent) +
  scale_fill_manual(values = c("male" = "skyblue", "female" = "pink")) +
  labs(title = "Predicted Probability of Good Credit Class by Gender",
       x = "Gender",
       y = "Predicted Probability",
       fill = "Gender") +
  theme_minimal()
```

Figure 2.3.21

Figure 2.3.22

This predicted probability bar chart, which is based on the regression model, clearly shows that women are less likely to be categorized as good credit risk in comparison to males.

Conclusion:

Thus, after conclusive evidence we reject the null hypothesis and accept the alternative hypothesis stating that:

An individual's gender being female increases the probability of them being classified as a bad credit risk.

Objective 2: To investigate the impact of employment duration on credit class classification

Employment duration is an important aspect of an individual's professional life that financial institutions look towards when classifying credit risk. A person's length of employment hints at their job stability which in turn strengthens their ability to repay loans.

Analysis Questions:

Analysis Question 1: What is the distribution of the ***employment*** column?

We begin by looking at the different values in the column ***employment*** and their frequency.

```
> data$employment = as.factor(data$employment)
> summary(data$employment)
      <1      >=7   1<=X<4   4<=X<7 unemployed
    1120     1227     2158     1199        296
```

Figure 2.3.23

>=7	1<=X<4	4<=X<7	<1	unemployed
1227	2158	1199	1120	296

Table 2.3.3

We now derive a new column with updated names for ***employment*** column and arrange the values for better visualization.

```
#Changing category names for more clarity
data$employment_categories = ifelse(data$employment == "<1", "Less than 1 year",
                                    ifelse(data$employment == ">=7", "7 or more years",
                                           ifelse(data$employment == "1<=X<4", "1 to 3 years",
                                                  ifelse(data$employment == "4<=X<7", "4 to 6 years",
                                                         ifelse(data$employment == "unemployed", "Unemployed", "NULL")))))
data$employment_categories = factor(data$employment_categories,
                                    levels = c("Unemployed",
                                               "Less than 1 year",
                                               "1 to 3 years",
                                               "4 to 6 years",
                                               "7 or more years"))
```

Figure 2.3.24

unemployed	->	Unemployed
<1	->	Less than 1 year
1<=X<4	->	1 to 3 years
4<=X<7	->	4 to 6 years
>=7	->	7 or more years

Table 2.3.4

To form a better understanding of the distribution of these values, we now create a pie chart:

```r
#Pie chart to show the distribution of employment duration
employment_summary = as.data.frame(table(data$employment_categories))
colnames(employment_summary) = c("employment categories","frequency")
employment_summary$proportions = (employment_summary$frequency / sum(employment_summary$frequency)) * 100
labels = paste0(employment_summary$employment_categories, " (", round(employment_summary$proportion, 1), "%)")
pie(
  employment_summary$frequency,
  labels = labels,
  col = c("#AEC6CF", "#77DD77", "#FFB7B2", "#C9A0DC", "#FFDAB9"),
  main = "Distribution of Employment Categories"
)
legend(
  "topright",
  legend = employment_summary$`employment categories`,
  fill = c("#AEC6CF", "#77DD77", "#FFB7B2", "#C9A0DC", "#FFDAB9"),
  title = "Employment Categories",
  cex = 0.4)
```

Figure 2.3.25

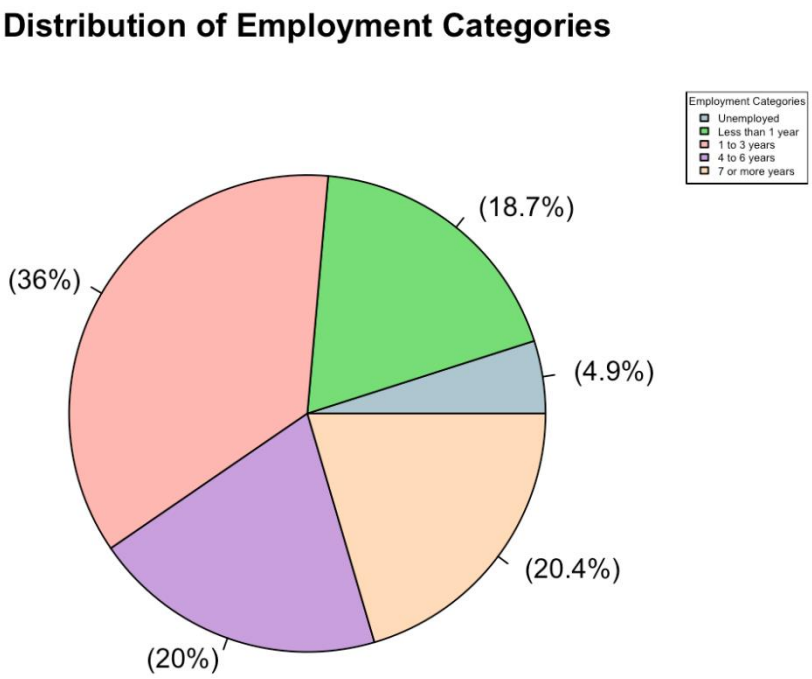

Figure 2.3.26

This pie chart shows us that the majority of the people represented in this dataset are the ones that have been employed for 1 to 3 years (36%), followed by 7 or more years (20.4%), then followed by 4 to 6 years which is just behind (20%), which is followed by those that have been employed for less than an year (18.7%) with the rest (4.9%) being unemployed.

The distribution of this data can be considered good if we were to do an analysis solely based on employment duration as the middle categories have a good spread. This data cannot

support analysis based on whether people are employed or unemployed based on the small amount of unemployed people in the dataset.

We can look at the credit risk classification for each category with a stacked bar chart to better understand the magnitude of the distribution while also taking a look at the credit class distribution.

```
#Frequency bar for employment categories
ggplot(data, aes(x = employment_categories, fill = class)) +
  geom_bar() +
  labs(title = "Frequency of Credit Ratings by Employemnt Category",
       x = "Employment Category",
       y = "Frequency",
       fill = "Credit Class") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Figure 2.3.27

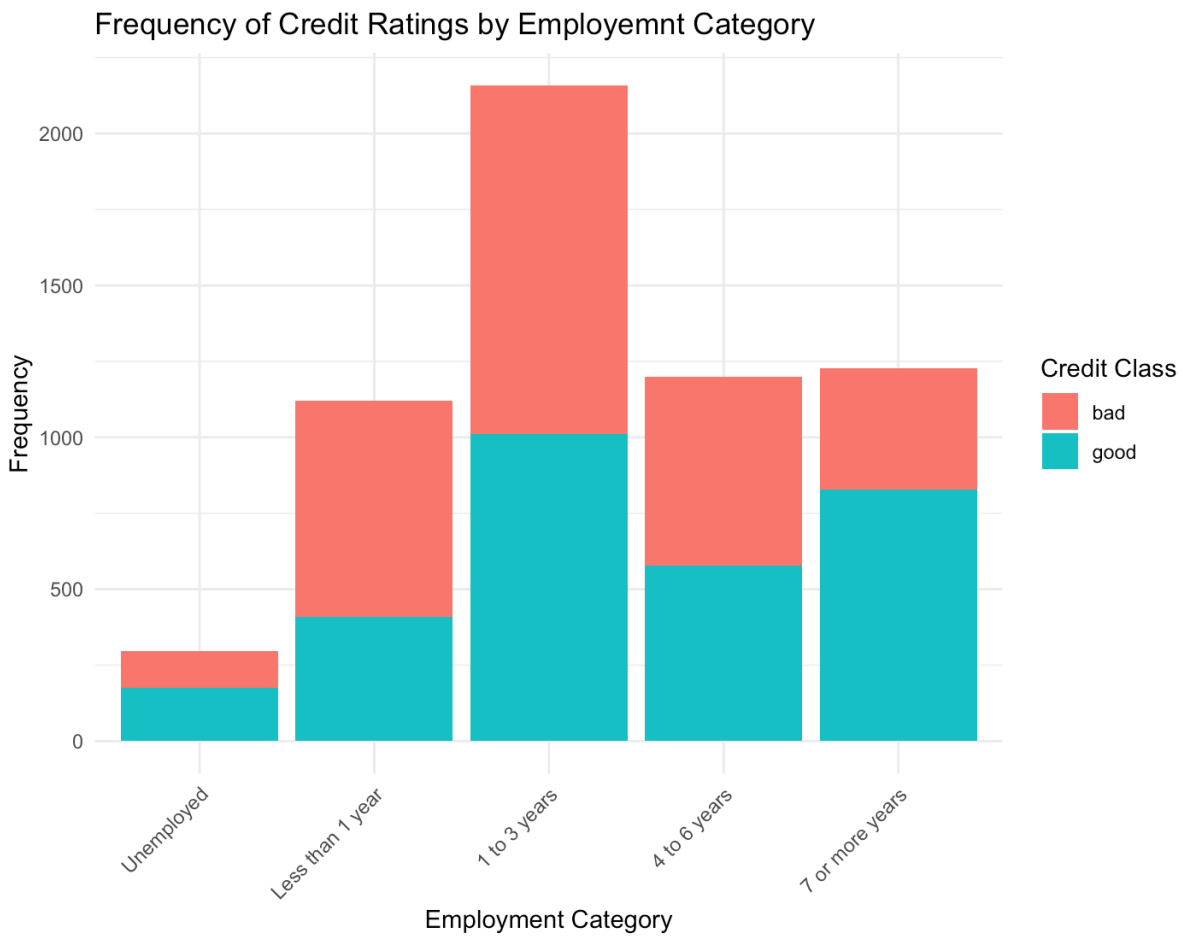

Figure 2.3.28

As we can see from both the charts people with employment duration 1 to 3 years are the most prevalent in terms of frequency while unemployed people are the least in frequency.

Analysis Question 2: Which factors within the ***employment*** column should be prioritized for analysis?

There are two ways to go about this dataset, the first approach is to divide the dataset into employed and unemployed categories while the other approach is to look at the employment duration of the people who are employed. We are going to follow the latter approach given the limited number of observations for the unemployed category and so that we can focus on a more meaningful analysis.

We will therefore filter our dataset to only include people who are employed, this results in us creating a new dataset.

```
#Filtering and excluding the unemployed values
employed_data = data %>%
  filter(employment_categories != "Unemployed")
employed_data$employment_categories = droplevels(employed_data$employment_categories,
                                                 employed_data$employment_categories == "Unemployed")
```

Figure 2.3.29

Analysis Question 3: What is the distribution of credit risk and what are some trends among the individuals who are employed?

We will now make a mosaic plot of the proportions of the dataset excluding the unemployed category to better understand the distribution of the dataset according to employment duration and their respective credit class.

```
#Mosaic plot to visualize the spread of employed people and credit class
library("ggmosaic")
ggplot(employed_data) +
  geom_mosaic(aes(x = product(employment_categories), fill = class), na.rm = TRUE) +
  labs(
    title = "Mosaic Plot of Credit Class by Employment Categories",
    x = "Employment Categories",
    y = "Proportion",
    fill = "Credit Class"
  ) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 0, hjust = 0.5),  # Adjust x-axis labels (no rotation)
    axis.title.y = element_text(margin = margin(r = 10))  # Add margin to y-axis title
  )
```

Figure 2.3.30

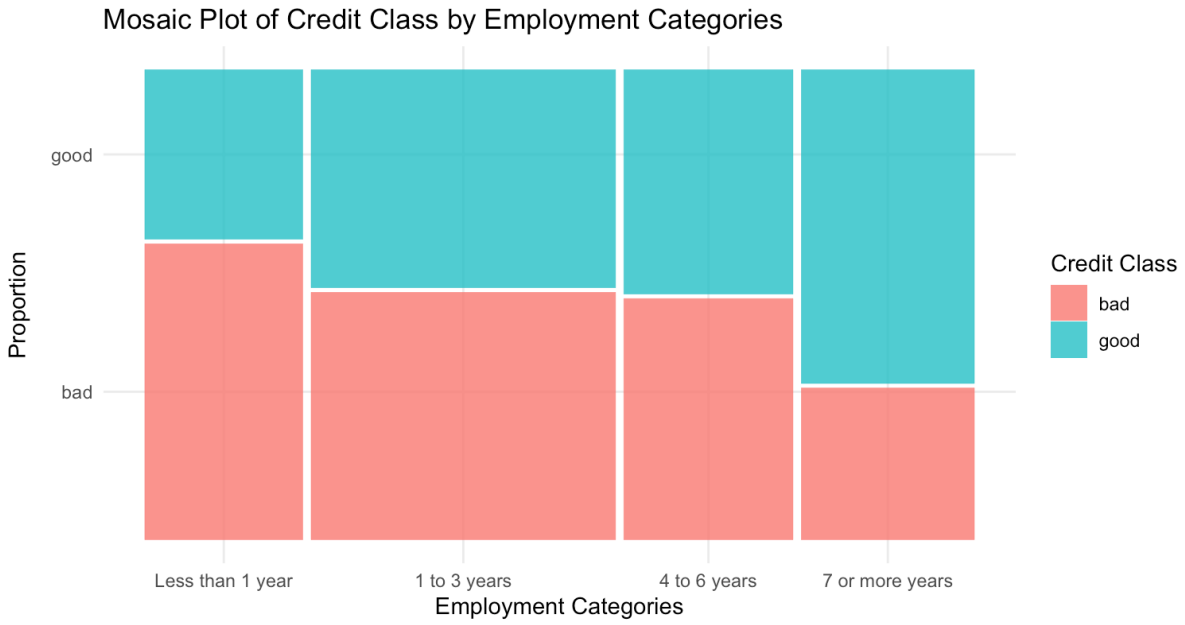

Figure 2.3.31

This mosaic plot is a very comprehensive way of visualizing our dataset. The width of the columns represents the proportion of individuals in each employment category while they are further divided into good and bad credit risk by the colors red or blue with red being bad credit risk and blue being good credit risk.

A decrease in bad credit risk looks evident with every increase in the employment duration category.

Literature Review:

Bruce Mclary, spokesman for the National Foundation for Credit Counseling, a Washington, D.C.-based non-profit organization affirms that the danger of job hopping is the inconsistency in a person's income and the difficulty it causes in keeping payments up to date with creditors. Mclary also stated that when vendors like to see consistency in employment to know that the loanee will be able to return the credit amount (George, 2016). Additionally, some loan types also have an impact on whether employment duration will be accounted for when classifying credit risk. According to research conducted by Maria George from Dublin Business School in 2016, most banks agree that employment stability has an impact of the chances of a loan application being accepted (George, 2016).

Putting all these factors together we come to an understanding that job instability, i.e. shorter employment durations, can have an adverse effect on a person's credit risk classification.

Hypothesis:

Null Hypothesis (H_0): The probability of being classified as a "bad" credit risk does not change with employment duration. Employment duration has no significant effect on credit classification.

Alternative Hypothesis (H_1): The probability of being classified as a "bad" credit risk decreases with an increase in employment duration.

Hypothesis Testing:

We begin our hypothesis testing my conducting a chi square test.

```
> employment_table = table(employed_data$class,employed_data$employment_categories)
> chisq.test(employment_table)

        Pearson's Chi-squared test

data:  employment_table
X-squared = 240.7, df = 3, p-value < 2.2e-16
```

Figure 2.3.32

The chi square test produces a p-value of less than 2.2e-16 which is highly significant thus indicating that employment categories have a highly significant impact on credit classification.

Now, we move to look at the strength of the relationship by calculating the Cramer's V value to quantify the strength of the relationship between employment categories and credit risk classification.

```
> assocstats(employment_table)
                    X^2 df P(> X^2)
Likelihood Ratio 244.88  3        0
Pearson          240.70  3        0

Phi-Coefficient   : NA
Contingency Coeff.: 0.201
Cramer's V        : 0.205
```

Figure 2.3.33

The test conducted above gives p values of nearly zero, which suggests ample evidence of rejecting the null hypothesis as the chances of the null hypothesis occurring are lower than a

value that the test can compute. The Cramer's V value of 0.205 signals a moderate relationship between the two columns.

We now further analyze the nature of this relationship by running a logistic regression model.

```
#Logistic Regression for employment categories
logistic_model_employed = glm(class_bin ~ employment_categories, family = "binomial", data = employed_data)
summary(logistic_model_employed)
```

Figure 2.3.34

```
Call:
glm(formula = class_bin ~ employment_categories, family = "binomial",
    data = employed_data)

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                          0.55681    0.06209   8.967  < 2e-16 ***
employment_categories1 to 3 years   -0.43060    0.07561  -5.695 1.23e-08 ***
employment_categories4 to 6 years   -0.48505    0.08483  -5.718 1.08e-08 ***
employment_categoriesMore than 7 years -1.28315    0.08698 -14.753  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7906.9  on 5703  degrees of freedom
Residual deviance: 7662.0  on 5700  degrees of freedom
AIC: 7670

Number of Fisher Scoring iterations: 4
```

Figure 2.3.35

The p value for each category is statistically significant indicating that all of the categories contribute independently to the prediction of credit risk classification.

We can observe that the value of the coefficient is positive at the intercept which represents the category of employment duration of less than 1 year and the coefficient value decreases as we increase the duration of employment.

We further convert the coefficients of the logistic regression into odds ratio to make understanding the data easier.

```
> odds = exp(coef(logisitic_model_employed))
> odds
                            (Intercept)      employment_categories1 to 3 years      employment_categories4 to 6 years
                              1.7450980                              0.6501184                              0.6156642
employment_categoriesMore than 7 years
                              0.2771626
```

Figure 2.3.36

The intercept category is our reference category “less than 1 year”. These odds calculate the likelihood of the category in focus being classified as bad credit risk. According to this, the odds of being classified as bad credit risk for people with an employment duration of less than 1 year are 1.74. The second category is people who have been employed for 1 to 3 years, and they show a clear change as the odds of being classified as bad credit risk jump down to 0.65. This downward trend continues in the next category of people with employment duration of 4 to 6 years having the odds of being classified as bad being 0.61. Finally in the last category of employment duration of more than 7 years, the trend reaches its lowest point with the odds reaching as low as 0.28.

We convert these odds into percentages to easier understand the probability of being classified as bad credit risk per category. We can then plot these percentages to get hold of the overall trend.

```
#Calculating probability of being bad credit risk per category
odds = exp(coef(logistic_model_employed))
Probability = odds / (1 + odds) * 100
bad_credit_data <- data.frame(
  Employment_Category = c("Less than 1 year", "1 to 3 years", "4 to 6 years", "More than 7 years"),
  Percentage_Bad = Probability
)

#Preparing data for bar chart
bad_credit_data$Employment_Category <- factor(
  bad_credit_data$Employment_Category,
  levels = c("Less than 1 year", "1 to 3 years", "4 to 6 years", "More than 7 years")
)

#bar chart to show chance of being bad credit risk per category
ggplot(bad_credit_data, aes(x = Employment_Category, y = Percentage_Bad, fill = Employment_Category)) +
  geom_bar(stat = "identity", width = 0.7) +
  labs(
    title = "Percentage of Bad Credit Risk by Employment Category",
    x = "Employment Categories",
    y = "Percentage of Bad Credit Risk (%)"
  ) +
  scale_fill_manual(values = c("red", "blue", "green", "purple")) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "none"
  )
```

Figure 2.3.37

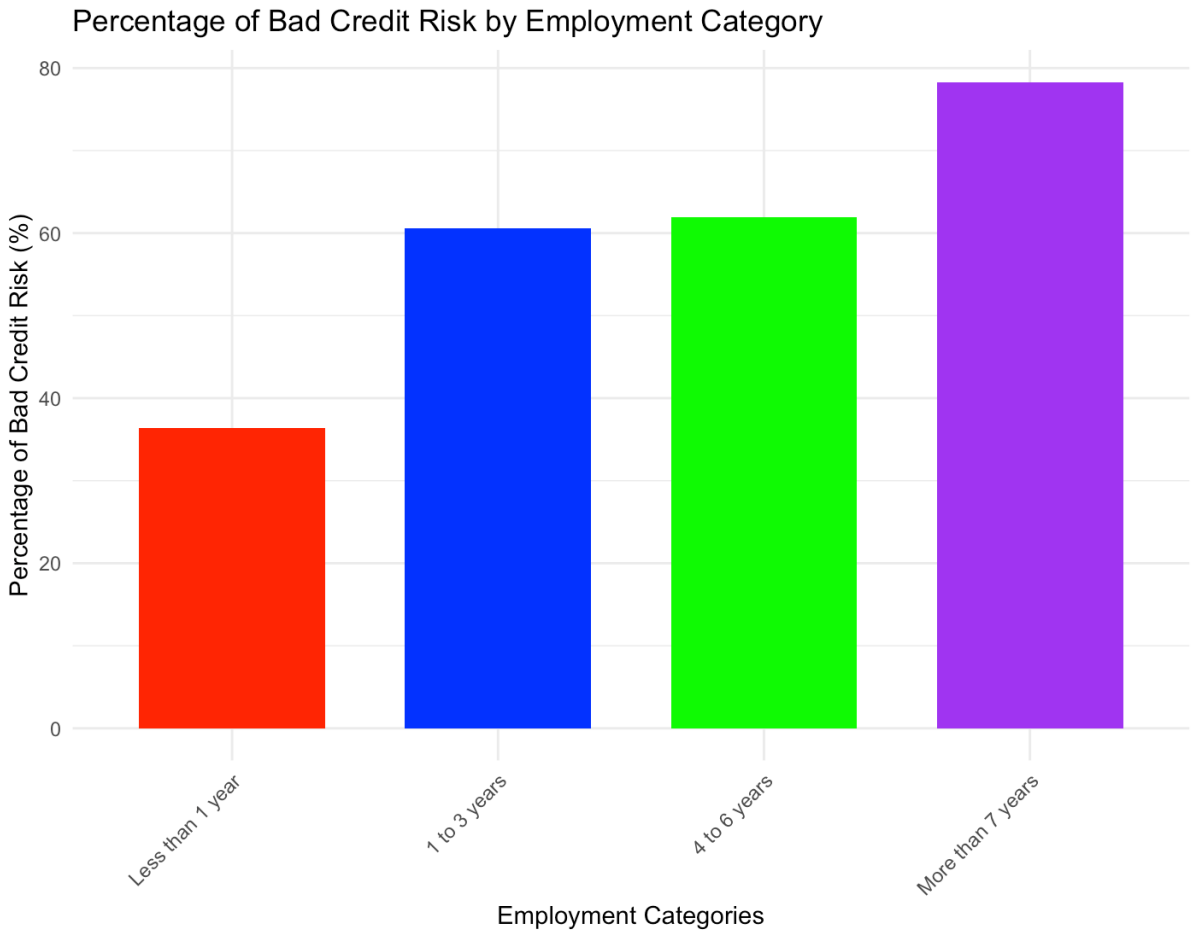

Figure 2.3.38

This graph clearly shows that the increase in employment duration directly corelates to a decrease in the probability of being classified as bad credit risk.

Conclusion:

Thus, we reject the null hypothesis and accept our alternative hypothesis which states that:

The probability of being classified as a "bad" credit risk decreases with an increase in employment duration.

Objective 3: To investigate the impact of *property_magnitude* and *num_dependents* on credit risk.

The property a person owns and the number of people dependent on him are some of the many factors that influence a lender's decision to classify a person as a good or bad credit

risk. Owning some sort of property signals financial stability, gives the lender reassurance of getting his money back, all be it through that property.

A high number of dependents also spark alarms for a lender as dependents are an incurring cost with no financial benefit, in other words they are a liability. Which strangles an individual's disposable income.

Analysis Questions:

Analysis Question 1: What is the distribution of the ***property_magnitude*** column?

The values present for the ***property_magnitude*** column in the dataset are:

```
> data$property_magnitude = as.factor(data$property_magnitude)
> summary(data$property_magnitude)
             car    life insurance no known property       real estate
            2156              1576              1002              1266
```

Figure 2.3.39

No known property	Cars	Real estate	Life insurance
1002	2156	1266	1576

Table 2.3.5

We can visualize the distribution of the types of property owned by individuals in our dataset using a pie chart.

```r
#Calculating probability of being bad credit risk per category
odds = exp(coef(logistic_model_employed))
Probability = odds / (1 + odds) * 100
bad_credit_data <- data.frame(
  Employment_Category = c("Less than 1 year", "1 to 3 years", "4 to 6 years", "More than 7 years"),
  Percentage_Bad = Probability
)

#Preparing data for bar chart
bad_credit_data$Employment_Category <- factor(
  bad_credit_data$Employment_Category,
  levels = c("Less than 1 year", "1 to 3 years", "4 to 6 years", "More than 7 years")
)

#bar chart to show chance of being bad credit risk per category
ggplot(bad_credit_data, aes(x = Employment_Category, y = Percentage_Bad, fill = Employment_Category)) +
  geom_bar(stat = "identity", width = 0.7) +
  labs(
    title = "Percentage of Bad Credit Risk by Employment Category",
    x = "Employment Categories",
    y = "Percentage of Bad Credit Risk (%)"
  ) +
  scale_fill_manual(values = c("red", "blue", "green", "purple")) +
  theme_minimal() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    legend.position = "none"
  )
```

Figure 2.3.40

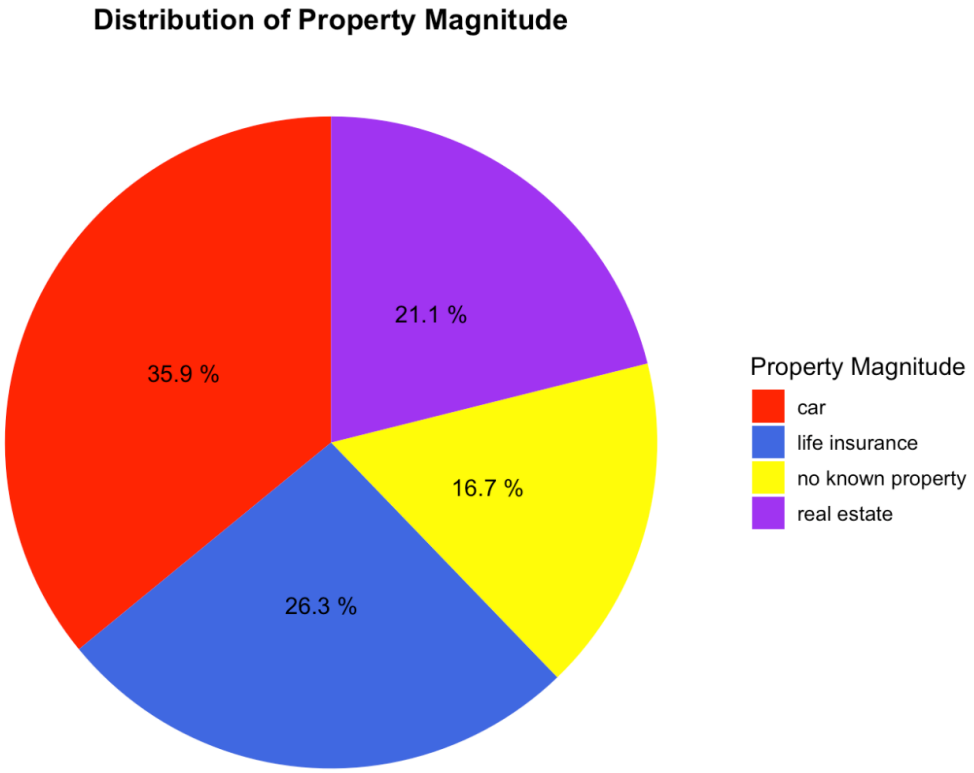

Figure 2.3.41

As we can observe from the pie chart above people owning cars form the majority comprising of 35.9% of the whole dataset, followed by people having life insurance who are a respectable 26.3%, people owning real estate are just behind being 21.3% and the least occurring category are the ones with no known property.

The data in this column is well spread, having sufficient values for each column this means that we will be able to build a good model based on our study objective.

We now build a frequency bar chart with good or bad credit classes to understand the relationship between property magnitude and credit class.

```
#frequency bar chart fro property magnitude
ggplot(data, aes(x = property_magnitude, fill = class)) +
  geom_bar(position = "dodge") +
  labs(title = "Count of Credit Classes by Property Magnitude",
       x = "Properrty Magnitude", y = "Count", fill = "Credit Class") +
  theme_minimal()
```

Figure 2.3.42

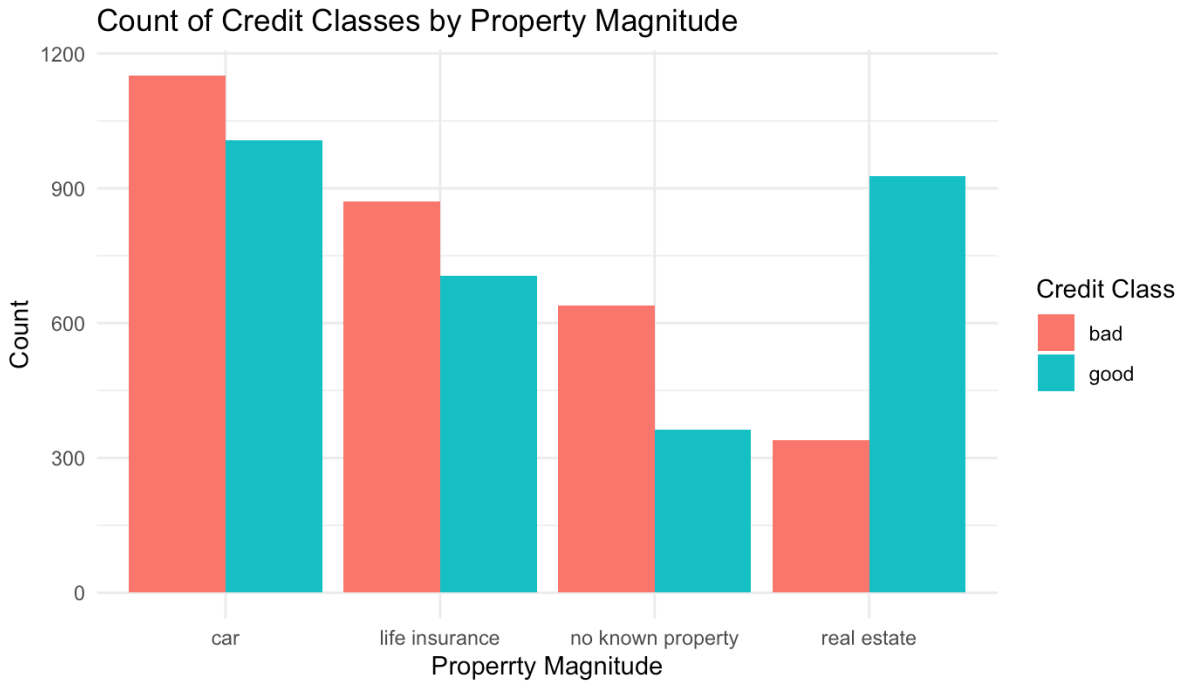

Figure 2.3.43

People who own real estate are the only category having more individuals with good credit risk, and it has almost 3x more individuals with good credit risk compared to individuals with bad credit risk. This shows that whether a person owns real estate is a major question that lenders consider.

Analysis Question 2: What is the distribution of the ***num_dependents*** column?

The ***num_dependents*** column consisted of various impossible values which had to undergo standardization. This cleaning resulted in the creation of a new ***new_num_dependents*** columns. (refer to Figure 1.50 for num_dependents cleaning)

The values present for the ***new_num_dependents*** column in the dataset are:

```
> summary(data$new_num_dependents)
   1    2
4827 1173
```

Figure 2.3.44

1	2
4827	1173

Table 2.3.6

We explore the distribution of these values using a pie chart:

```
#Number of dependents pie chart
dependents_counts <- table(data$new_num_dependents)
percentages <- round(prop.table(dependents_counts) * 100, 1)
labels <- paste(percentages, "%", "(", as.numeric(dependents_counts), ")")
colors <- c("royalblue", "brown")
pie3D(dependents_counts,
      labels = labels,
      explode = 0,
      col = colors,
      radius = 0.8,
      labelcex = 1.0,
      theta = 1.2)
title(main = "Distribution of Number of Dependents", cex.main = 1.5)
legend("topright",
       legend = names(dependents_counts),
       fill = colors,
       title = "Number of Dependents",
       cex = 0.8)
```

Figure 2.3.45

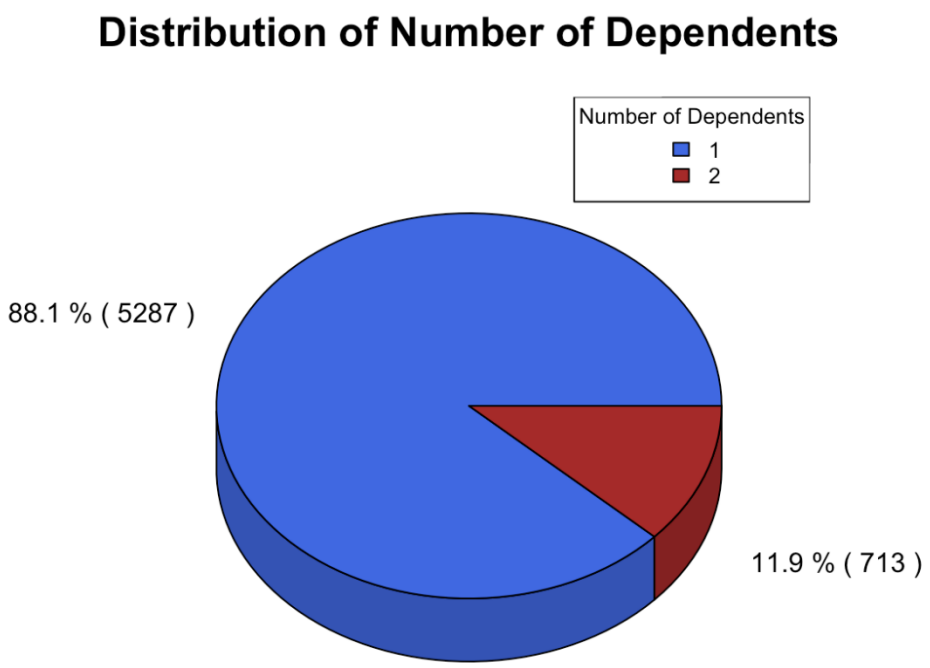

Figure 2.3.46

The dataset's heavily skewed distribution, with 80.4% having one dependent and only 19.6% having two, reduces the reliability of credit risk predictions for individuals with two dependents in comparison to those with one dependent. The model's training is dominated by the one-dependent group, which could lead to less accurate or biased predictions when applied to the under-represented two-dependent group.

We use a bar chart to showcase the frequency of the distribution with good or bad credit classification.

```
#Frequency bar chart for number of dependents
ggplot(data, aes(x = new_num_dependents, fill = class)) +
  geom_bar(position = "dodge") +
  labs(title = "Count of Credit Classes by Number of Dependents",
       x = "Number of Dependents", y = "Count", fill = "Credit Class") +
  theme_minimal()
```

Figure 2.3.47

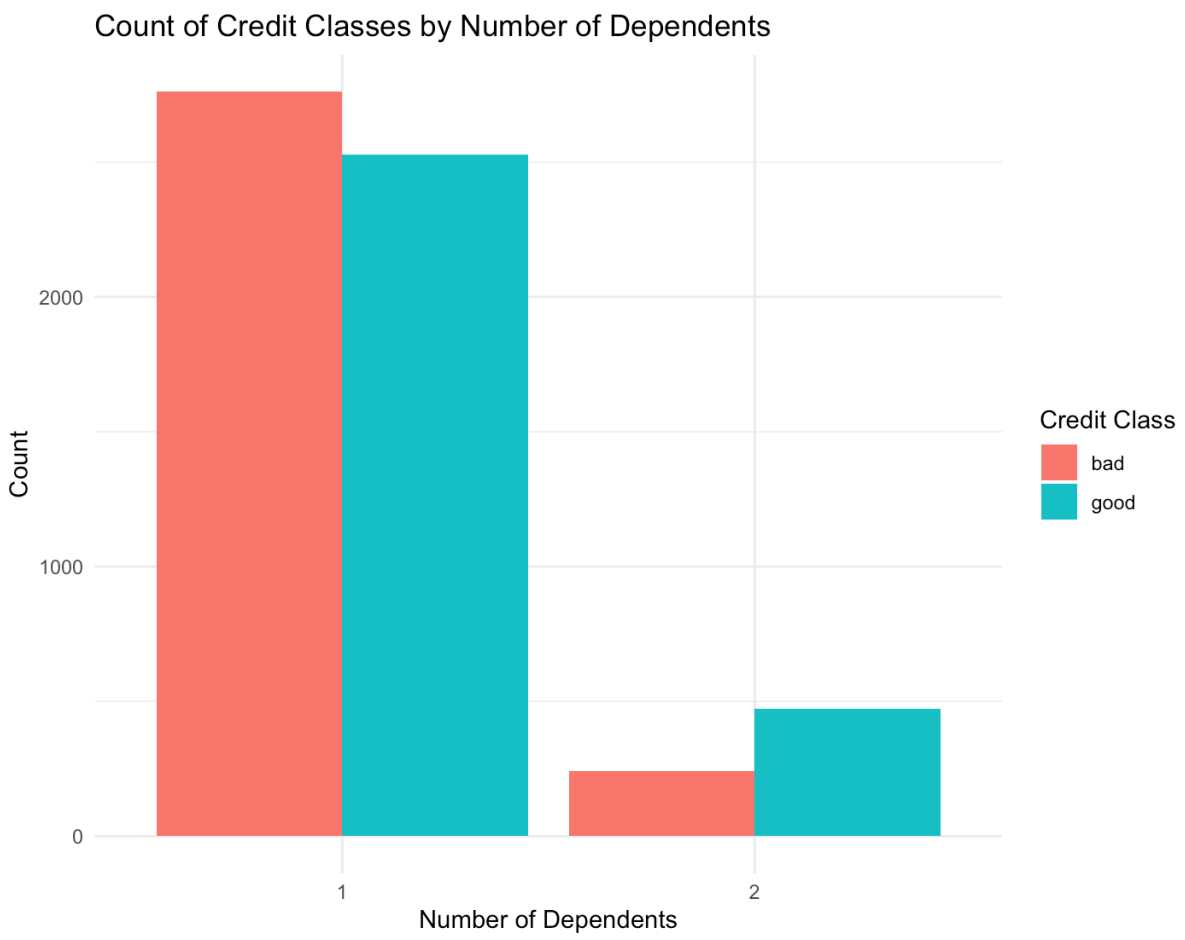

Figure 2.3.48

We can clearly observe discrepancy in the number of individuals with one and the number of individuals with two dependents, this is in line with the previous pie chart. An important detail we observe hair is that people with one dependent have around 500 more entries of good credit class in comparison to bad credit class. However, people with two dependents have around 500 more entries for bad credit class compared to good credit class.

Analysis Question 3: How does an individual's property magnitude and number of dependents interact with their credit risk classification?

To get an idea of the good and bad credit ratings for each category we will make use of a grouped bar chart:

```r
#grouped bar chart for property magnitude and number of dependents by credit class
ggplot(data, aes(x = property_magnitude, fill = class)) +
  geom_bar(position = "fill") +  # Fill makes it proportionate
  facet_wrap(~ new_num_dependents) +  # Separate panels by no_of_dependents
  labs(title = "Distribution of Credit Class by Property Magnitude and Dependents",
       x = "Property Magnitude",
       y = "Proportion",
       fill = "Credit Class") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_flip()
```

Figure 2.3.49

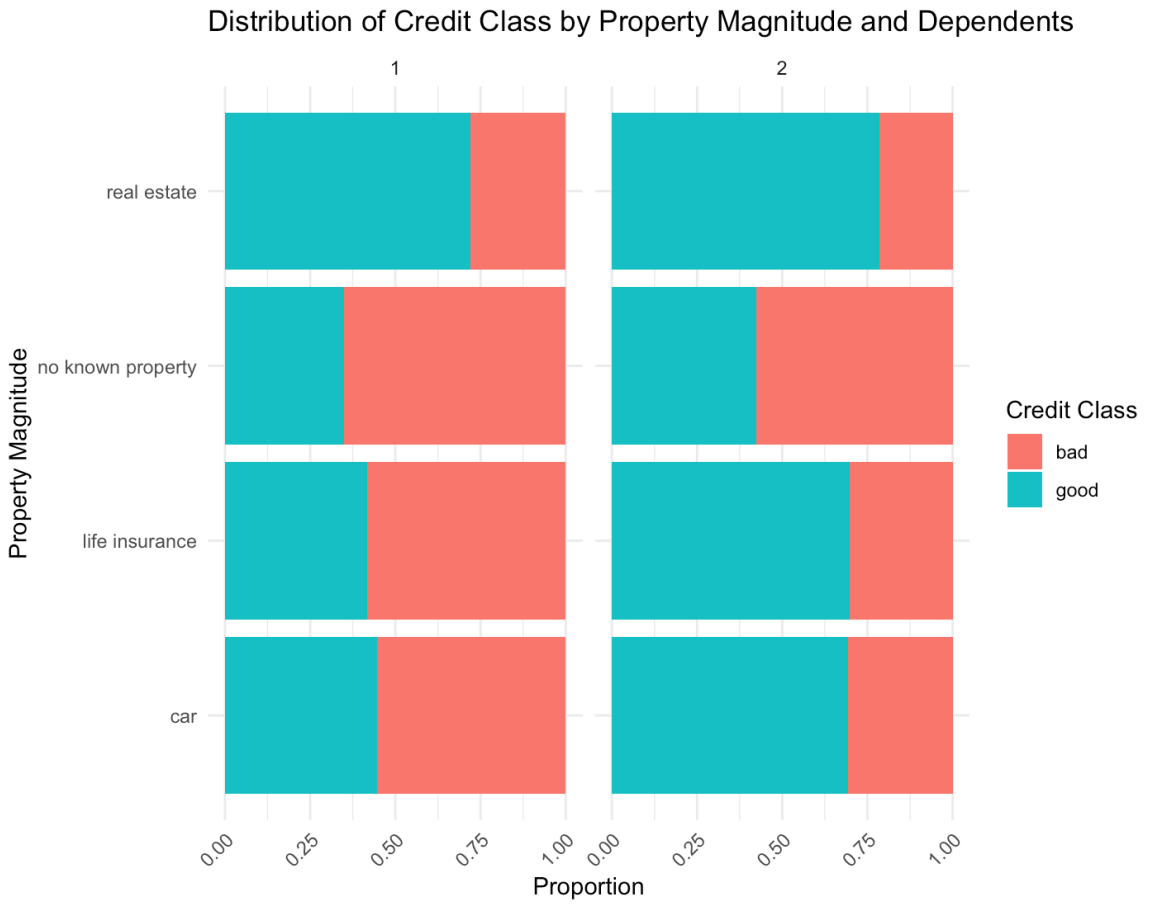

Figure 2.3.50

Across both groups of dependents, individuals with no known property have the highest proportion of bad credit risk, with the effect exacerbated for those with two dependents. People who own real estate have the highest proportion of good credit risk in both dependent groups, and this proportion further increases within the two-dependent group. For all other property categories, the proportion of bad credit risk increases for individuals with two dependents compared to those with one.

Literature Review:

Credit risk assessment hinges on financial stability, with income being a key predictor. Higher income typically correlates with higher FICO scores, as individuals with more disposable income are better positioned to manage debt (Arya et al., 2011). Increased financial strain may hinder debt management, increasing default risk and negatively affecting credit classification.

Real estate ownership signals financial stability and is viewed favorably by financial institutions. For example, the Bank-Fund Federal Credit Union demonstrates this preference by offering significantly larger loans against property compared to unsecured loans or those secured by other assets (Bogetic, 2005). This favorable treatment underscores the perceived lower risk associated with property owners, who are seen as more financially reliable. The lack of such a tangible asset may therefore negatively impact an individual's perceived creditworthiness.

However, the financial burden of dependents significantly impacts disposable income, regardless of real estate ownership. The USDA estimates raising a child to age 18 costs a middle-income family $233,610, and thus reduces disposable income (Edwards, 2017). This effect makes a case that increase in dependents decreases a person's financial stability.

Hypothesis:

Null Hypothesis (H_0): There is no significant relationship between the combination of owning real estate and the number of dependents on an individual's credit class.

Alternative Hypothesis (H_1): Individuals who own real estate and have a lower number of dependents have better odds of being classified as "good" credit risk compared to those who have no known property and have a higher number of dependents.

Hypothesis Testing:

We begin our hypothesis testing by performing a chi-square test on both the number of dependents and property magnitude against credit class.

```
> table_dependents = table(data$new_num_dependents, data$class)
> chisq.test(table_dependents)

        Pearson's Chi-squared test with Yates' continuity correction

data:  table_dependents
X-squared = 54.124, df = 1, p-value = 1.882e-13
```

Figure 2.3.51

First for the chi test between number of dependents and credit class we get a p-value of 1.882e-13 which indicates a highly significant relationship.

```
> table_property = table(data$property_magnitude, data$class)
> chisq.test(table_property)

        Pearson's Chi-squared test

data:  table_property
X-squared = 377.33, df = 3, p-value < 2.2e-16
```

Figure 2.3.52

Next, for the chi test between property magnitude and credit class we get a p-value of less than 2.2e-16 which is again highly significant and even more significant than the previous chi test.

The chi tests provide major evidence against the null hypothesis and indicate that there is a relationship between the variables in question. To understand and quantify the strength of these relationships, we proceed to perform Cramer's v test on the variables.

```r
#Cleaning data for cramer's v matrix
library(vcd)
library(gtools)
library(reshape2)
cramers_data <- data.frame(
  data$property_magnitude,
  data$new_num_dependents,
  data$class
)
cramers_v_matrix <- function(cramers_data) {
  vars <- colnames(cramers_data)
  n <- length(vars)
  matrix <- matrix(NA, ncol = n, nrow = n)
  colnames(matrix) <- rownames(matrix) <- vars
  for (i in 1:n) {
    for (j in 1:n) {
      if (i == j) {
        matrix[i, j] <- 1
      } else {
        tbl <- table(cramers_data[[vars[i]]], cramers_data[[vars[j]]])
        cramer_v <- assocstats(tbl)$cramer
        matrix[i, j] <- cramer_v
      }
    }
  }
  return(matrix)
}
```

Figure 2.3.53

```r
cramer_matrix <- cramers_v_matrix(cramers_data)
cramer_df <- melt(cramer_matrix)
colnames(cramer_df) <- c("Variable1", "Variable2", "CramersV")

#Creating a heatmap to visualize cramer's v values
ggplot(cramer_df, aes(x = Variable1, y = Variable2, fill = CramersV)) +
  geom_tile(color = "white") +  # Add tile borders
  scale_fill_gradient2(low = "white", high = "blue", midpoint = 0.2, name = "Cramér's V") +
  geom_text(aes(label = round(CramersV, 2)), color = "black", size = 4) +
  theme_light() +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1, size = 10),
    axis.text.y = element_text(size = 10),
    panel.grid = element_blank()
  ) +
  labs(
    title = "Cramér's V Matrix Heatmap",
    x = "Variables",
    y = "Variables"
  )
```

Figure 2.3.54

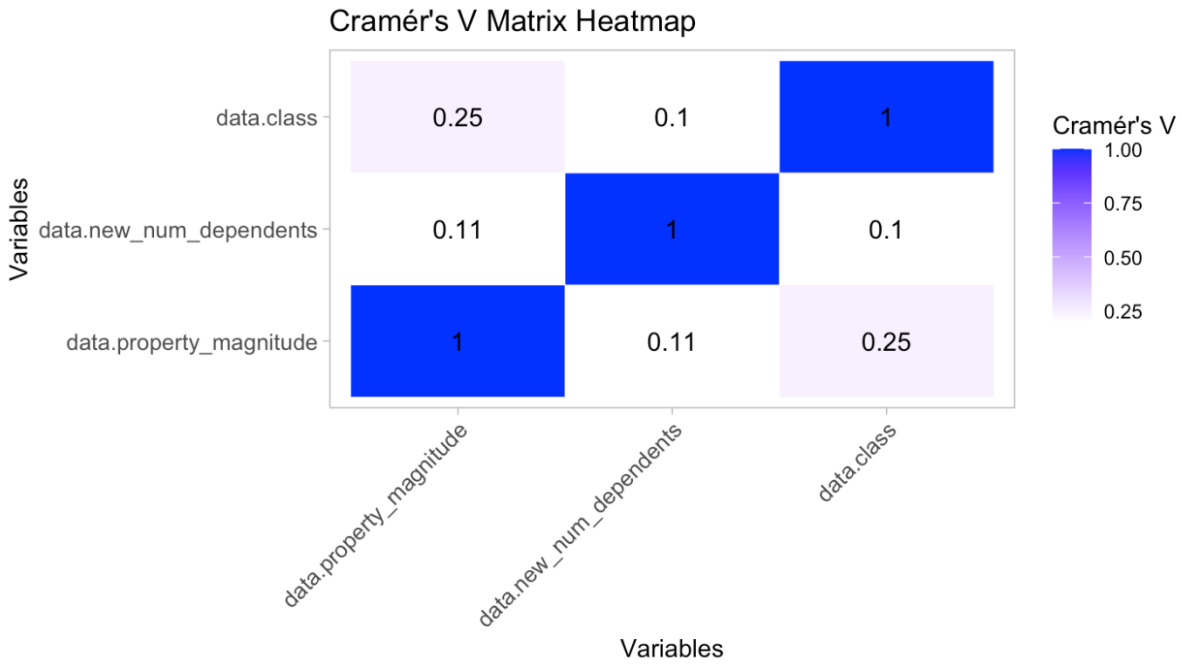

Figure 2.3.55

The heatmap above visually showcases the strength of the association between all the variables in question using the Cramer's V value. Since a darker color indicates a stronger relationship, we can observe that the strongest relationship from those relevant to us are between property and class. The property and class columns have a moderate relationship between each other while the number of dependents and class have a weak relationship between them.

As we are now aware of the significance of the categories and their effect on class. It is now time to build a model in order to reach a decisive conclusion.

We now build a logistic regression model to conduct further analysis.

```
> model <- glm(class_bin ~ new_num_dependents * property_magnitude,family = binomial, data = data)
> summary(model)

Call:
glm(formula = class_bin ~ new_num_dependents * property_magnitude,
    family = binomial, data = data)

Coefficients:
                                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                                            -0.068842   0.046767  -1.472 0.141018
new_num_dependents2                                    -0.440343   0.123734  -3.559 0.000373 ***
property_magnitudelife insurance                        0.001548   0.073975   0.021 0.983300
property_magnitudeno known property                    -0.311536   0.088741  -3.511 0.000447 ***
property_magnitudereal estate                           1.029644   0.083395  12.347  < 2e-16 ***
new_num_dependents2:property_magnitudelife insurance   -0.223570   0.177102  -1.262 0.206812
new_num_dependents2:property_magnitudeno known property -0.322343   0.202436  -1.592 0.111313
new_num_dependents2:property_magnitudereal estate       0.717619   0.215413   3.331 0.000864 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8317.8  on 5999  degrees of freedom
Residual deviance: 7860.3  on 5992  degrees of freedom
AIC: 7876.3

Number of Fisher Scoring iterations: 4
```

Figure 2.3.56

According to our hypothesis, we are only concerned with the category which has one dependent and owns real estate. The category relevant to us has the lowest p value out of all the categories in the logistic regression. This shows that this is the most significant in predicting credit class. Additionally, the coefficient for this category is also the highest, indicating that this category improves the chances of being classified as “good” credit risk by the most margins compared to all other categories.

We will now plot these log odds so that we can visually assess our findings.

```r
#Cleaning model to plot odds
library(broom)
model_summary <- tidy(model, conf.int = TRUE, exponentiate = TRUE)

#odds plot for visualization
ggplot(model_summary, aes(x = term, y = estimate, ymin = conf.low, ymax = conf.high)) +
  geom_point() +
  geom_errorbar(width = 0.2) +
  scale_y_log10() +
  labs(
    title = "Odds Ratios for Model Coefficients",
    x = "Terms",
    y = "Odds Ratio (log scale)"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Figure 2.3.57

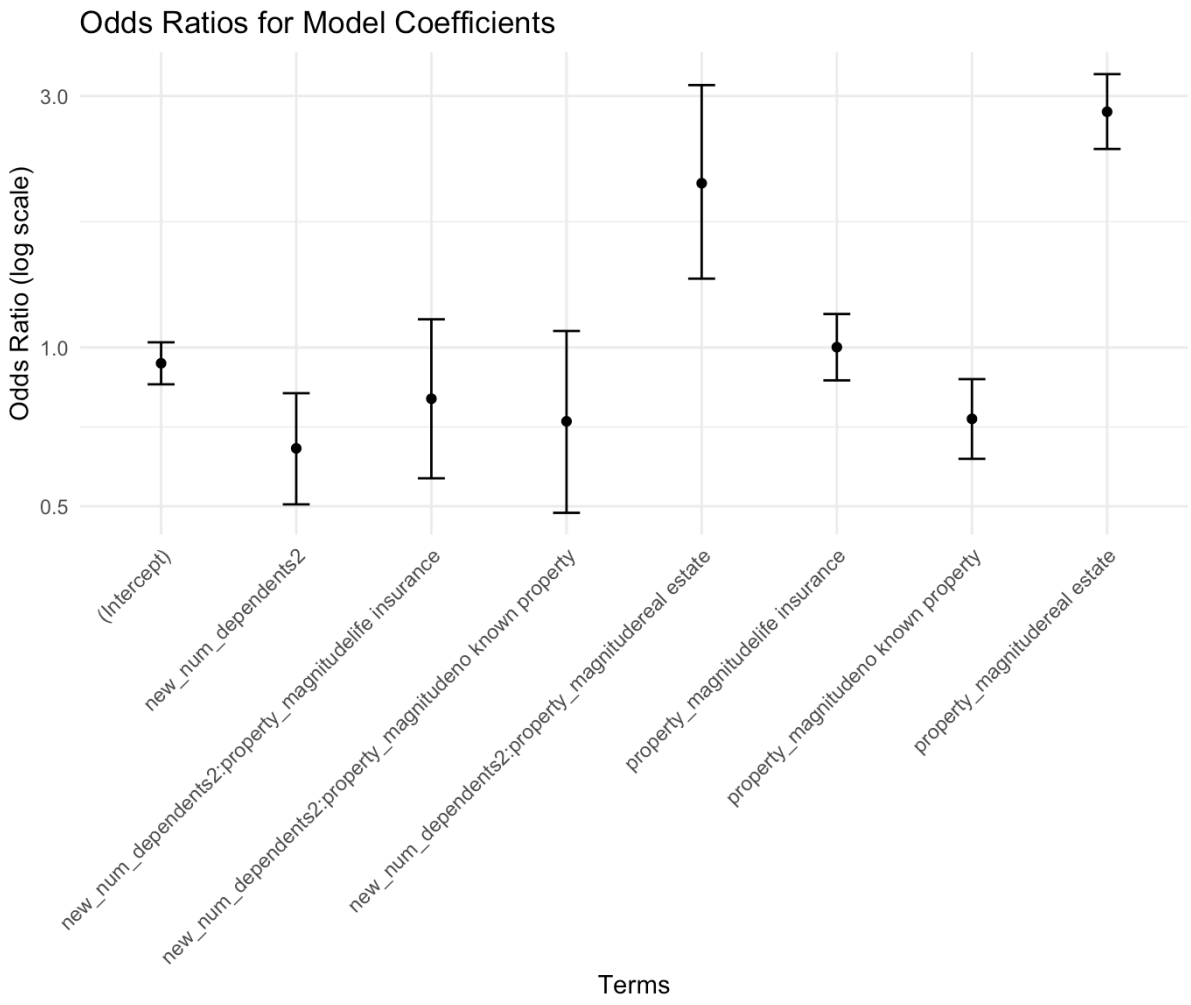

Figure 2.3.58

The graph above illustrates the odds of being classified as having good credit risk for each category. The intercept represents individuals with one dependent. The highest odds of good credit risk classification are observed among individuals who own real estate and have one dependent. Conversely, individuals with no known property and two dependents have significantly lower odds of good credit risk classification compared to those who own real estate and have one dependent.

These findings are perfectly in line with our alternative hypothesis.

Conclusion:

We will reject the null hypothesis and accept the alternative hypothesis stating that: Individuals who own real estate and have a lower number of dependents have better odds of being classified as "good" credit risk compared to those who have no known property and have a higher number of dependents.

3.2 Group Hypothesis & Testing

Hypothesis:

Null Hypothesis (H_0): Bad credit risk classification is not significantly influenced by whether a person is female, who neither owns real estate nor owns a phone and is seeking a relatively higher amount of credit. The combination of these factors do not significantly affect the likelihood of being classified as a bad credit risk.

Alternative Hypothesis (H_1): Bad credit risk classification is significantly influenced by whether a person is female, who neither owns real estate nor owns a phone, and is seeking a relatively higher amount of credit. The combination of these factors significantly increases the odds of being classified as a bad credit risk.

Group Hypothesis Testing:

We begin our hypothesis testing by conducting Chi-Square tests for all the variables with credit class:

```
> #Preparing data for Chi Square tests
> table_real_estate <- table(credit_risk$owns_real_estate, credit_risk$class_binary)
> table_owntele <- table(credit_risk$owntele, credit_risk$class_binary)
> table_credit_amount <- table(credit_risk$credit_amount_cat, credit_risk$class_binary)
> table_gender <- table(credit_risk$gender, credit_risk$class_binary)
>
> #Performing Chi square tests
> chi_test_real_estate <- chisq.test(table_real_estate)
> chi_test_owntele <- chisq.test(table_owntele)
> chi_test_credit_amount <- chisq.test(table_credit_amount)
> chi_test_gender <- chisq.test(table_gender)
>
> #Summarising Chi Square Findings
> chi_summary <- data.frame(
+   Variables = c("Real Estate","Own Telephone", "Credit Amount", "Gender"),
+   P_Value = c(chi_test_real_estate$p.value, chi_test_owntele$p.value, chi_test_credit_amount$p.value, chi_test_gender$p.value)
+ )
> print(chi_summary)
      Variables      P_Value
1   Real Estate 5.311961e-77
2 Own Telephone 3.086852e-40
3 Credit Amount 1.397774e-06
4        Gender 1.752971e-06
```

Figure 3.2.1

Chi-square tests are used to find out if there is an association between two categorical variables. In this case, the test evaluates the likelihood that there is no association between

each of the predictor variables and the target variable, "Credit Class." Specifically, it is calculating the probability that the predictor variable has no effect on the target variable.

The results from the Chi Square test are:

Variables	p-value
Real Estate	5.311961e-77
Own Telephone	3.086852e-40
Credit Amount	1.397774e-06
Gender	1.752971e-06

Table 3.2.1

We observe that the variable with the lowest p-value is Real Estate, its p-value shows overwhelming evidence of real estate influencing credit risk classification. Next, is Own Telephone with a p-value which once again shows overwhelming evidence of influencing credit risk classification. The other two variables of credit amount and gender have much larger p-values compared to the first two variables, but these too are way smaller than the significance threshold. Thus, all four variables in our hypothesis significantly influence credit risk classification, albeit some more than others.

The sole job of the chi-square test is to inform us that an association exists between the variables, the practical effect or strength of this association is not in the scope of the chi test. This is why we will calculate the Cramer's V value of our variables to assess the practical effect they have on credit class.

```
> #Performing cramer's V
> library(vcd)
> cramer_real_estate <- assocstats(table_real_estate)
> cramer_owntele <- assocstats(table_owntele)
> cramer_credit_amount <- assocstats(table_credit_amount)
> cramer_gender <- assocstats(table_gender)
>
> #Summarizing cramer's V findings
> cramer_summary <- data.frame(
+   Variable = c("owns_real_estate", "owntele", "credit_amount_cat", "gender"),
+   Cramer_V = c(cramer_real_estate$cramer, cramer_owntele$cramer, cramer_credit_amount$cramer, cramer_gender$cramer)
+ )
> print(cramer_summary)
          Variable   Cramer_V
1  owns_real_estate 0.24018526
2           owntele 0.17177688
3 credit_amount_cat 0.06263319
4            gender 0.06206621
```

Figure 3.2.2

The results from the Cramer's V are:

Variables	Cramer's V
Real Estate	0.24018526
Own Telephone	0.17177688
Credit Amount	0.06263319
Gender	0.06206621

Table 3.2.2

The Cramér's V values for our variables align with the findings of the chi-square test. Although the Cramér's V values are generally low, only "Real Estate" demonstrates a medium effect, while the other variables exhibit a weak effect on credit class. These values are consistent with the trend observed in the p-values from the chi-square test, where the variable with the highest association also shows the largest effect, as expected.

To visualize and better understand the extent to which each variable influences the credit class rating, we will create a radar chart.

```
> #Preparing data for radar chart
> library(fmsb)
> cramer_data <- cramer_summary$Cramer_V
> radar_data <- as.data.frame(t(cramer_data))
> colnames(radar_data) <- cramer_summary$Variable
>
> #Setting radar chart boundaries
> radar_data <- rbind(rep(1, nrow(cramer_summary)),
+                     rep(0, nrow(cramer_summary)),
+                     radar_data)
>
> radarchart(
+   radar_data,
+   axistype = 1,
+   cglcol = "gray",
+   cglty = 1,
+   axislabcol = "transparent",
+   pcol = "blue",
+   pfcol = scales::alpha("blue", 0.5),
+   plwd = 2,
+   title = "Cramér's V Radar Chart"
+ )
>
> legend(
+   "topright",
+   legend = paste0(cramer_summary$Variable, ": ", round(cramer_summary$Cramer_V, 2)),
and values
+   col = "black",
+   bty = "n",
+   cex = 0.8
+ )
```

Figure 3.2.3

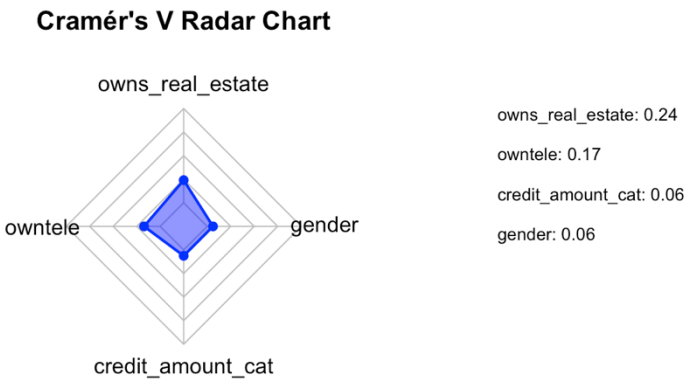

Figure 3.2.4

The edges of this radar chart represent a Cramér's V value of 1, which signifies a perfect association. In this graph, the further a point is from the centre, the greater the effect that variable has on influencing the credit class. A variable with no effect would be located at the centre. From the perspective of our hypothesis, this graph demonstrates that all variables have some level of effect on credit classification.

We will now create an Artificial Neural Network to check if gender, owning or not owning a telephone, owning or not owning real estate and credit amount are significant predictors of credit risk classification.

This means that if our model has a higher accuracy than random guessing then our chosen variables will be proven to be significant predictors of credit class.

```
#Preparing data for NN model.
library(nnet)
library(caret)
library(dplyr)
selected_vars = credit_risk[, c("class_binary", "owns_real_estate", "owntele", "credit_amount_cat", "gender")]
dummy_vars = dummyVars("~ .", data = selected_vars[, -which(names(selected_vars) %in% c("class_binary"))])
encoded_vars = as.data.frame(predict(dummy_vars, selected_vars))
encoded_vars$class_binary = selected_vars$class_binary

set.seed(123)
trainIndex = createDataPartition(encoded_vars$class_binary, p = 0.8, list = FALSE)
trainData = encoded_vars[trainIndex, ]
testData = encoded_vars[-trainIndex, ]

set.seed(123)
nn_model <- nnet(class_binary ~ ., data = trainData, size = 4, decay = 0.1, maxit = 500, linout = FALSE)
```

Figure 3.2.5

The above section of the code prepares the dataset to be used in the Neural Network model. This includes converting categorical columns into a format the model can understand. After that, the data is split, with 80% used for training the model and 20% saved for testing how well the model performs.

As can be seen in the code, we have built a neural network model consisting of 8 input nodes because of the number predictor columns that were created after one-hot encoding, a hidden layer with 4 neurons, and a single output node for binary classification. Each input node is linked to every neuron in the hidden layer and vice versa. This way the network is able to capture complex relationships between the input variables and predict the likelihood of being classified as a "good" or "bad" credit risk.

Below is a visual representation of the model.

```
> library(igraph)
> #Define nodes
> input_nodes <- paste0("i", 1:8)  # Input layer (8 nodes)
> hidden_nodes <- paste0("h", 1:5)  # Hidden layer (5 nodes)
> output_nodes <- "o"  # Output layer (1 node)
> #Combine nodes
> all_nodes <- c(input_nodes, hidden_nodes, output_nodes)
> #Define edges
> edges <- c(
+   # Input to Hidden
+   c("i1", "h1"), c("i2", "h1"), c("i3", "h1"), c("i4", "h1"), c("i5", "h1"), c("i6", "h1"), c("i7", "h1"), c("i8", "h1"),
+   c("i1", "h2"), c("i2", "h2"), c("i3", "h2"), c("i4", "h2"), c("i5", "h2"), c("i6", "h2"), c("i7", "h2"), c("i8", "h2"),
+   c("i1", "h3"), c("i2", "h3"), c("i3", "h3"), c("i4", "h3"), c("i5", "h3"), c("i6", "h3"), c("i7", "h3"), c("i8", "h3"),
+   c("i1", "h4"), c("i2", "h4"), c("i3", "h4"), c("i4", "h4"), c("i5", "h4"), c("i6", "h4"), c("i7", "h4"), c("i8", "h4"),
+   c("i1", "h5"), c("i2", "h5"), c("i3", "h5"), c("i4", "h5"), c("i5", "h5"), c("i6", "h5"), c("i7", "h5"), c("i8", "h5"),
+   # Hidden to Output
+   c("h1", "o"), c("h2", "o"), c("h3", "o"), c("h4", "o"), c("h5", "o")
+ )
> #Convert edges to matrix
> edges_matrix <- matrix(edges, ncol = 2, byrow = TRUE)
> #Plot the graph
> nn_graph <- graph_from_edgelist(edges_matrix, directed = TRUE)
> #Custom positions
> layout_positions <- matrix(NA, nrow = length(all_nodes), ncol = 2)
> #Layer positions
> layout_positions[1:8, ] <- cbind(1, seq(-8, -1, length.out = 8))
> #Hidden layer
> layout_positions[9:13, ] <- cbind(3, seq(-6, -2, length.out = 5))
> #Output layer
> layout_positions[14, ] <- cbind(5, -4)
> #Plot neural network
> plot(
+   nn_graph,
+   layout = layout_positions,
+   vertex.color = "lightblue",  # Node color
+   vertex.size = 30,  # Node size
+   vertex.label.cex = 0.8,  # Label size
+   edge.arrow.size = 1.0,  # Larger arrows for better clarity
+   edge.color = "gray",  # Edge color
+   main = "Neural Network Representation"
+ )
```

Figure 3.2.6

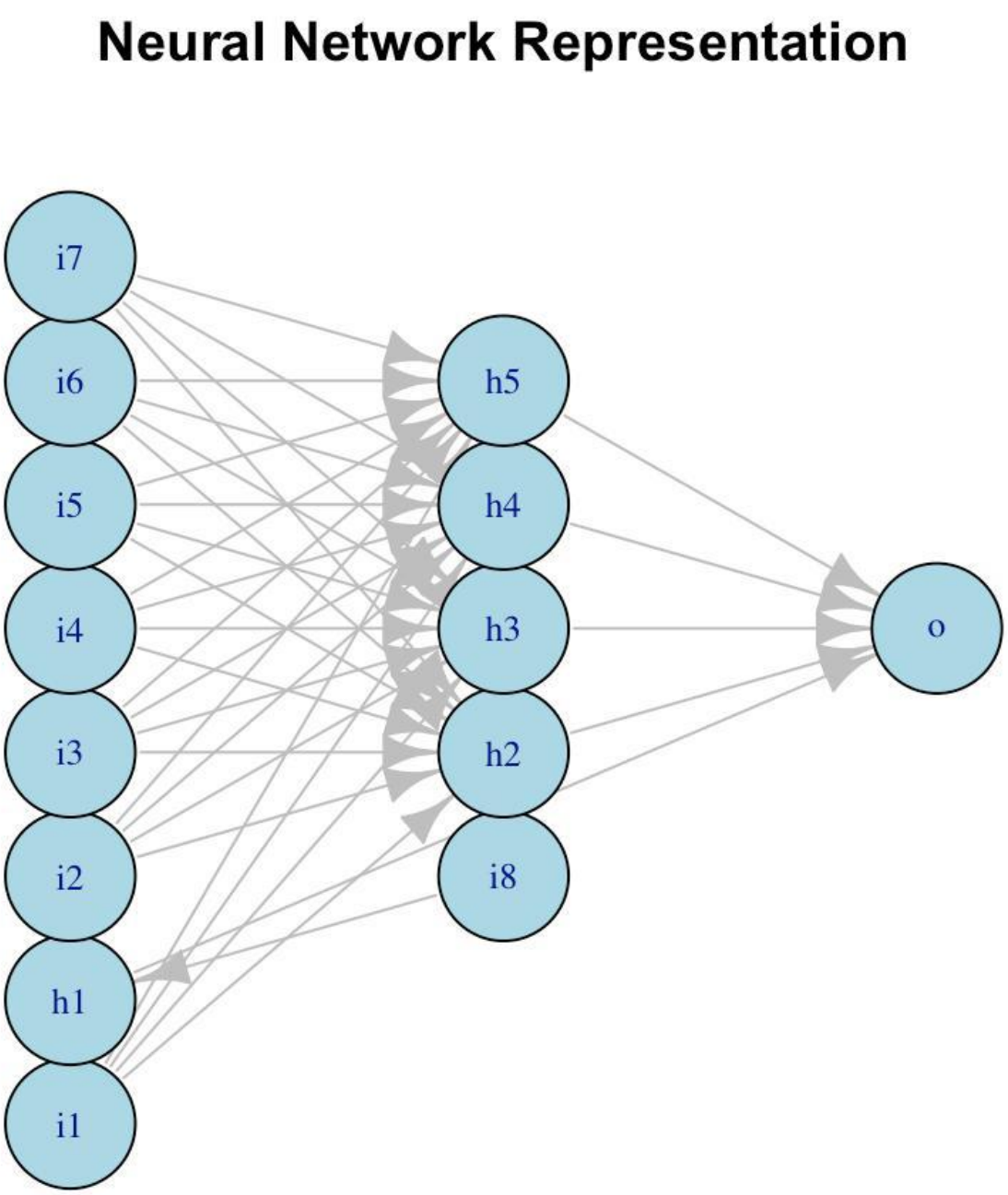

Figure 3.2.7

The chart represents an 8-5-1 neural network where the input layer from i1 to i8 passes information about the variables such as gender, real estate ownership, phone ownership and credit amount to the hidden layer from h1 to h5. The hidden layer processes combinations of these inputs to identify patterns that influence the classification of credit risk. The final output layer (o) produces a prediction such as whether a customer is classified as a bad credit risk. This structure allows the network to analyse the complex relationships among input variables and assess their collective impact on credit risk classification. For example, input node i1

might represent gender distinguishing between male and female. While, i2 could indicate real estate ownership and i8 might denote the credit amount categories such as high or low. These variables feed into the hidden nodes, where their interactions are processed to inform the output.

This neural network is directly relevant to the hypothesis Bad credit risk classification is significantly influenced by whether a person is female, who neither owns real estate nor owns a phone, and is seeking a relatively higher amount of credit. Each input node corresponding to these features contributes to the patterns learned in the hidden layer. For example, hidden node h1 might detect that being female i1 , not owning real estate i2 and seeking a high credit amount i8 collectively increase the likelihood of being classified as a bad credit risk. Other hidden nodes, such as h3 and h4 could capture additional interactions involving variables like phone ownership i3 or savings status. The output node aggregates these learned patterns and produces a final prediction based on the combined influence of all input variables.

In relation to the hypothesis, the network effectively evaluates the relationships between variables. The inclusion of nodes like (gender), (real estate ownership), (phone ownership), and (credit amount) ensures the model considers all mentioned in the hypothesis. If these have significant weights in the network's computations, it supports the hypothesis that being female, not owning real estate or a phone, and seeking a high credit amount contribute to bad credit risk classification. The model's structure allows it to learn these patterns and network's learned weights and predictions quantify the of each factor.

We now evaluate the accuracy of the model by generating predictions and testing them against the data that was set aside earlier for validation.

```
> predictions <- predict(nn_model, testData[, -which(names(testData) == "class_binary")], type = "raw")
> predictions <- predictions[, 1]
> predicted_classes <- ifelse(predictions > 0.5, 1, 0)
>
> confusionMatrix(as.factor(predicted_classes), as.factor(testData$class_binary))
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 446 226
         1 154 374

               Accuracy : 0.6833
                 95% CI : (0.6562, 0.7096)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3667

 Mcnemar's Test P-Value : 0.0002703

            Sensitivity : 0.7433
            Specificity : 0.6233
         Pos Pred Value : 0.6637
         Neg Pred Value : 0.7083
             Prevalence : 0.5000
         Detection Rate : 0.3717
   Detection Prevalence : 0.5600
      Balanced Accuracy : 0.6833

       'Positive' Class : 0
```

Figure 3.2.8

The model achieves an accuracy ranging from 65.62% to 70.96%, as indicated by the 95% confidence interval, demonstrating that it performs significantly better than random guessing. This confirms that the selected variables, such as gender, real estate ownership, phone ownership, and credit amount, play a meaningful role in predicting credit risk. The accuracy, combined with the balanced performance metrics, highlights that these variables are significant predictors of credit classification.

Since we are now certain of our variables' effects on credit class, it is finally time to assess what impact these variables have on credit risk classification. Specifically, we will now determine whether the categories mentioned in the hypothesis for these variables cause the credit risk to be more tilted toward a "bad" or "good" classification. For this task, we will make use of a logistic regression model.

```
> finalmodel = glm(class_binary ~ owns_real_estate * owntele * credit_amount_cat * gender ,
+                  data = credit_risk, family = "binomial")
> summary(finalmodel)

Call:
glm(formula = class_binary ~ owns_real_estate * owntele * credit_amount_cat *
    gender, family = "binomial", data = credit_risk)

Coefficients:
                                                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                                                        1.61774    0.22345   7.240  4.5e-13 ***
owns_real_estateNo                                                -0.57521    0.24945  -2.306 0.021117 *
owntelenone                                                       -0.51650    0.24581  -2.101 0.035621 *
credit_amount_cathigh                                             -0.02511    0.38757  -0.065 0.948351
genderfemale                                                       0.88370    0.51572   1.714 0.086617 .
owns_real_estateNo:owntelenone                                    -1.03131    0.27641  -3.731 0.000191 ***
owns_real_estateNo:credit_amount_cathigh                          -0.98843    0.40907  -2.416 0.015681 *
owntelenone:credit_amount_cathigh                                  0.63867    0.47567   1.343 0.179376
owns_real_estateNo:genderfemale                                   -1.06658    0.54671  -1.951 0.051069 .
owntelenone:genderfemale                                          -1.89652    0.54088  -3.506 0.000454 ***
credit_amount_cathigh:genderfemale                                -2.35855    0.69592  -3.389 0.000701 ***
owns_real_estateNo:owntelenone:credit_amount_cathigh               0.28971    0.50239   0.577 0.564167
owns_real_estateNo:owntelenone:genderfemale                        1.90559    0.57894   3.292 0.000996 ***
owns_real_estateNo:credit_amount_cathigh:genderfemale              2.30354    0.73509   3.134 0.001726 **
owntelenone:credit_amount_cathigh:genderfemale                     2.79155    0.80614   3.463 0.000534 ***
owns_real_estateNo:owntelenone:credit_amount_cathigh:genderfemale -2.69950    0.85768  -3.147 0.001647 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8317.8  on 5999  degrees of freedom
Residual deviance: 7537.1  on 5984  degrees of freedom
AIC: 7569.1

Number of Fisher Scoring iterations: 4
```

Figure 3.2.9

The logistic regression model above calculates the log odds to be classified as “good” credit risk based on the various categories from the variables in focus. The last coefficient on the logistic regression is the one that is relevant to our hypothesis.

The p-value is less than 0.05 thus the interaction is significant. The coefficient –2.69950 signifies that the likelihood of being classified as “good” credit risk decreases considerably. This is in line with our hypothesis.

The model above includes all inter-variable effects and interactions to capture potential complex patterns. However, not all interactions have a significant impact on credit class. To assess whether this complexity is necessary, we will compare the full model with a reduced model, which includes only the main effects of the variables. By performing ANOVA, we can determine if the added complexity of the full model provides a significant improvement in explaining credit class compared to the simpler reduced model.

```
> reduced_model <- glm(class_binary ~ owns_real_estate + owntele + credit_amount_cat + gender,
+                      data = credit_risk, family = "binomial")
> anova(reduced_model, finalmodel, test = "Chisq")
Analysis of Deviance Table

Model 1: class_binary ~ owns_real_estate + owntele + credit_amount_cat +
    gender
Model 2: class_binary ~ owns_real_estate * owntele * credit_amount_cat *
    gender
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1      5995     7667.5
2      5984     7537.1 11   130.33 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 3.2.10

The ANOVA test compares the reduced model we just created to the original model. This comparison yields a highly significant result, with a p-value of less than 2.2e-16, which is far below the 0.001 threshold. The result shows a 130.33-point decline in deviance, indicating that the full model, with all the interactions, provides a significantly better fit. Thus, it stands clear that the complexity of the full model is necessary to better explain credit class classification.

The coefficient log-odds from the full model have been converted into odds to enhance interpretability and facilitate a clearer understanding of their impact.

```
#Calculating odds of bad credit risk
odds <- exp(coef(finalmodel))
odds_df <- data.frame(
  Term = names(odds),
  Odds = odds
)
odds_df <- odds_df[order(-odds_df$Odds), ]

#plotting odds
ggplot(odds_df, aes(x = reorder(Term, Odds), y = Odds)) +
  geom_segment(aes(x = Term, xend = Term, y = 0, yend = Odds), color = "black") +
  geom_point(size = 4, color = "red") +
  coord_flip() +
  labs(
    title = "Lollipop Chart of Odds from Logistic Regression Model",
    x = "Predictors and Interactions",
    y = "Odds"
  ) +
  theme_minimal()
```

Figure 3.2.11

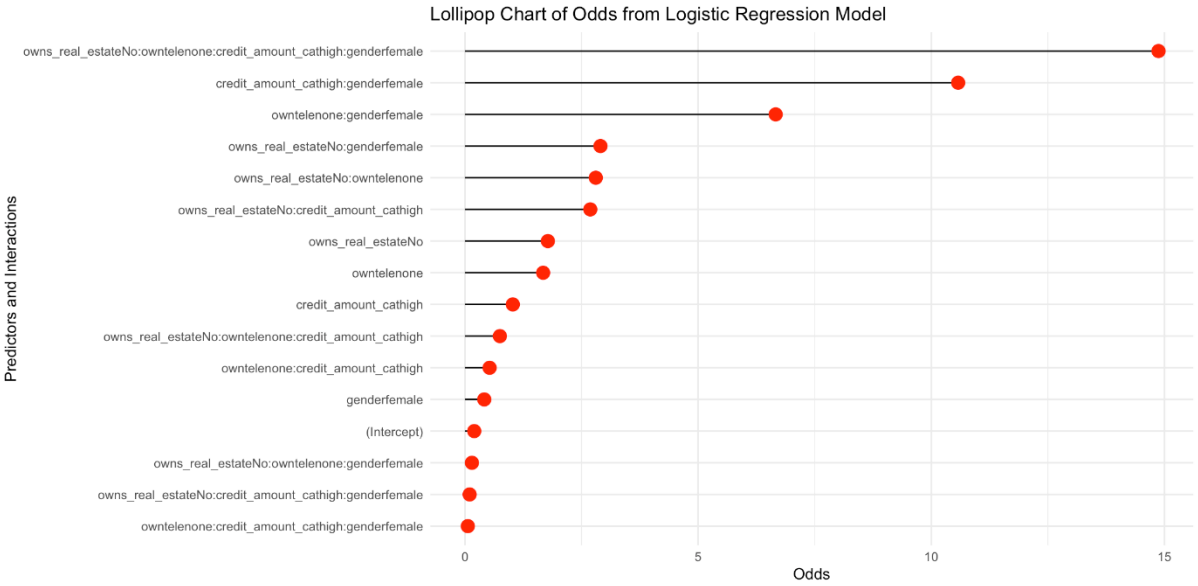

Figure 3.2.12

The chart above illustrates the odds of being classified as a “bad” credit risk based on the logistic regression model. The category relevant to our hypothesis is the first one in the chart. It is evident that the selected categories significantly increase the likelihood of being classified as a “bad” credit risk, as indicated by the substantial exponential increase in odds.

Conclusion:

According to the results from the above performed tests, we will confidently reject our null hypothesis and accept our alternative hypothesis which states that:

Bad credit risk classification is significantly influenced by whether a person is female, who neither owns real estate nor owns a phone, and is seeking a relatively higher amount of credit. The combination of these factors significantly increases the odds of being classified as a bad credit risk.