Project

Jani Shariff Shaik

2024-04-24

INITIAL PART

Step 1: Data Import and Exploration

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.2
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
## Loading required package: lattice
```

```
library(glmnet)
```

```
## Warning: package 'glmnet' was built under R version 4.3.3
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-8
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.3.2
```

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

```
##
## Attaching package: 'pROC'
```

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

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

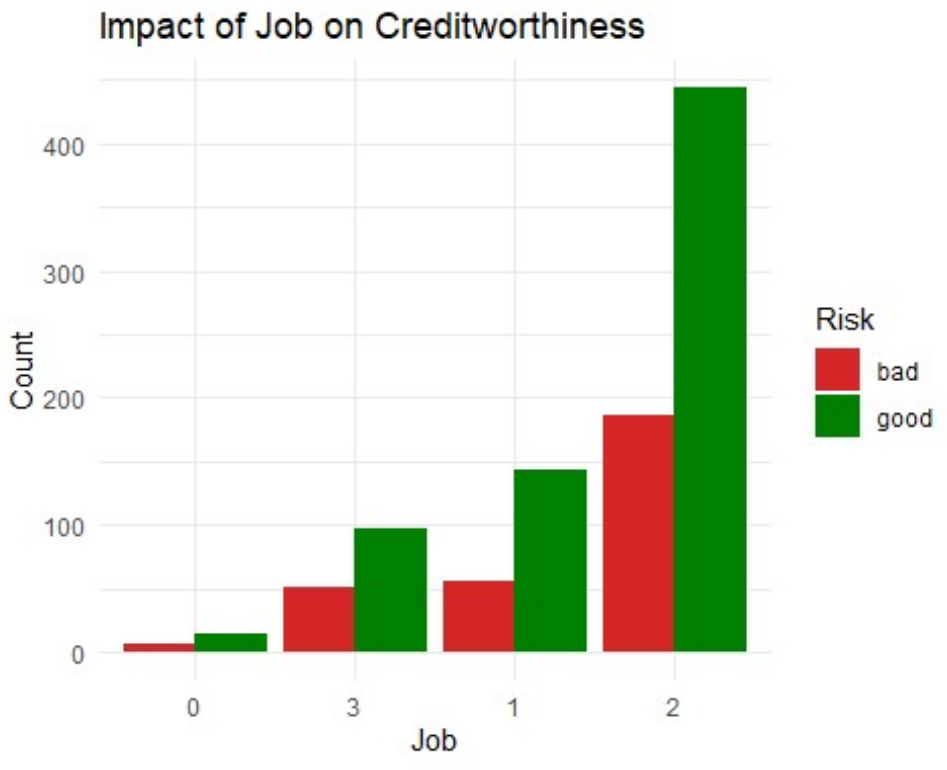
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

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

# Load the dataset (replace 'credit_data.csv' with your dataset file name)
data <- read.csv("C:/Users/janis/OneDrive/Desktop/german_credit_data.csv")

# Explore the structure and summary statistics of the dataset
str(data)

## 'data.frame':    1000 obs. of  11 variables:
##  $ X               : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Age             : int  67 22 49 45 53 35 53 35 61 28 ...
##  $ Sex             : chr  "male" "female" "male" "male" ...
##  $ Job             : int  2 2 1 2 2 1 2 3 1 3 ...
##  $ Housing         : chr  "own" "own" "own" "free" ...
##  $ Saving.accounts : chr  NA "little" "little" "little" ...
##  $ Checking.account: chr  "little" "moderate" NA "little" ...
##  $ Credit.amount   : int  1169 5951 2096 7882 4870 9055 2835 6948 3059
5234 ...
##  $ Duration        : int  6 48 12 42 24 36 24 36 12 30 ...
##  $ Purpose         : chr  "radio/TV" "radio/TV" "education"
"furniture/equipment" ...
##  $ Risk            : chr  "good" "bad" "good" "good" ...

# Handle missing values
# Total number of null values in each column
null_counts <- colSums(is.na(data))
null_counts

##                X              Age              Sex              Job
##                0                0                0                0
##          Housing  Saving.accounts Checking.account    Credit.amount
##                0              183              394                0
##         Duration          Purpose             Risk
##                0                0                0

data <- mutate_all(data, ~ ifelse(is.na(.), 'Not Available', .))
```

Question 1: What is the impact of employment status (job) on creditworthiness?

```
library(dplyr)
library(ggplot2)

# Count frequencies of each job category
job_counts <- table(data$Job)

# Find job categories with non-zero counts
non_zero_jobs <- names(job_counts[job_counts > 0])

# Filter data to include only job categories with non-zero counts
```

```r
data_filtered <- data[data$Job %in% non_zero_jobs, ]

# Reorder levels of "Job" based on count
data_filtered$Job <- factor(data_filtered$Job, levels =
names(sort(table(data_filtered$Job))))

# Plot the reordered and filtered data with grouped bars
ggplot(data_filtered, aes(x = Job, fill = Risk)) +
  geom_bar(position = "dodge", stat = "count") +  # Use position = "dodge"
for grouped bars
  labs(title = "Impact of Job on Creditworthiness", x = "Job", y = "Count") +
  scale_fill_manual(values = c("good" = "#008000", "bad" = "#d62728")) +
  theme_minimal()
```



Statistical analysis

```r
# Perform Chi-square test for Job and Risk
chi_square_job <- chisq.test(table(data_filtered$Job, data_filtered$Risk))

# Print the results
print("Chi-square test for Job and Risk:")

## [1] "Chi-square test for Job and Risk:"

print(chi_square_job)

##
##  Pearson's Chi-squared test
##
```

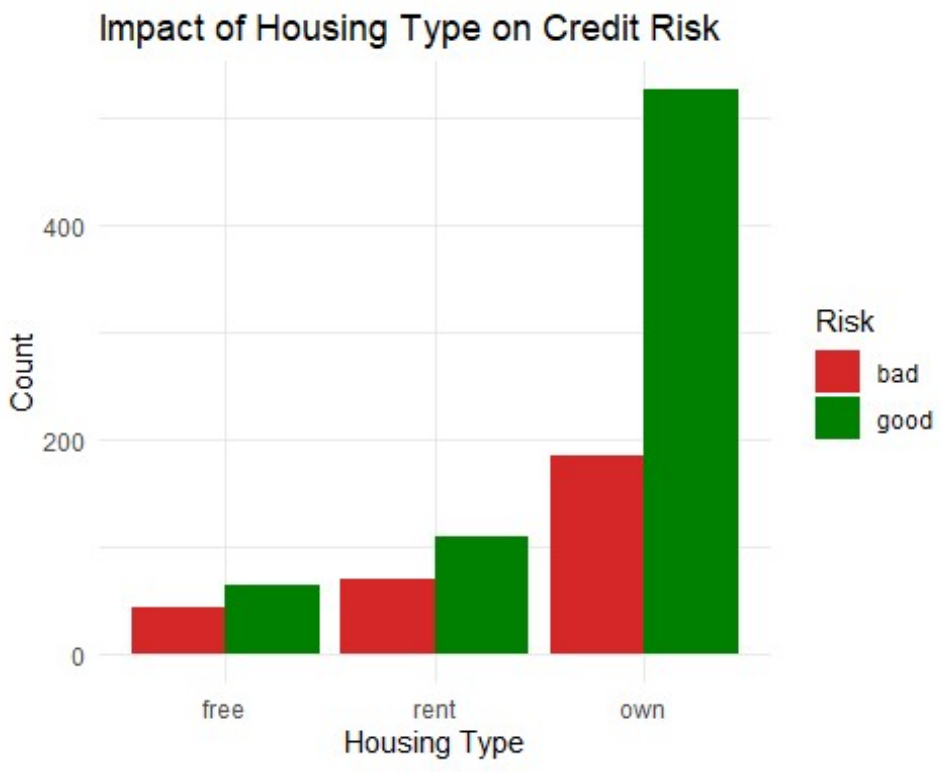
```
## data:  table(data_filtered$Job, data_filtered$Risk)
## X-squared = 1.8852, df = 3, p-value = 0.5966
```

The Chi-square test results indicate that there is no significant association between no of jobs and creditworthiness (Risk) (χ^2 = 1.8852, df = 3, p = 0.5966, α = 0.05). Thus, we fail to reject the null hypothesis, suggesting that job category does not have a significant impact on credit risk.

Question 2 :

Does the type of housing (own, rent, or free) affect the likelihood of good credit risk?

```
# Reorder levels of "Housing" based on count
data$Housing <- factor(data$Housing, levels =
names(sort(table(data$Housing))))

# Plot the reordered data
ggplot(data, aes(x = Housing, fill = Risk)) +
  geom_bar(position = "dodge") +  # Grouped bar plot
  labs(title = "Impact of Housing Type on Credit Risk",
       x = "Housing Type", y = "Count") +
  scale_fill_manual(values = c("good" = "#008000", "bad" = "#d62728")) +  #
Color definition
  theme_minimal()
```



Statistical testing

```
# Create contingency table for Housing and Risk
housing_contingency <- table(data$Housing, data$Risk)
```

```r
# Print contingency table
print("Contingency table for Housing and Risk:")
```

```
## [1] "Contingency table for Housing and Risk:"
```

```r
print(housing_contingency)
```

```
##
##        bad good
##   free  44   64
##   rent  70  109
##   own  186  527
```

```r
# Perform Chi-square test for Housing and Risk
chi_square_housing <- chisq.test(housing_contingency)
print("Chi-square test for Housing and Risk:")
```

```
## [1] "Chi-square test for Housing and Risk:"
```

```r
print(chi_square_housing)
```

```
##
##  Pearson's Chi-squared test
##
## data:  housing_contingency
## X-squared = 18.2, df = 2, p-value = 0.0001117
```

Contingency Table: The contingency table shows the frequencies of "bad" and "good" credit risks for each housing type (free, rent, own). Chi-square Test Result: The chi-square test yielded a test statistic of 18.2 with 2 degrees of freedom and a p-value of 0.0001117. Inference: The low p-value (< 0.05) indicates that there is a significant association between housing type and credit risk. Therefore, housing type may be a relevant factor in determining credit risk.

Question 3:

How do different levels of saving and checking accounts (e.g., little, moderate, rich) relate to credit risk?

```r
# Load required libraries
library(dplyr)
library(ggplot2)

# Create contingency table for Saving.accounts and Risk
saving_contingency <- table(data$Saving.accounts, data$Risk)

# Print contingency table
print("Contingency table for Saving.accounts and Risk:")
```

```
## [1] "Contingency table for Saving.accounts and Risk:"
```

```r
print(saving_contingency)
```

```
##
##                 bad good
##   little        217  386
##   moderate       34   69
##   Not Available  32  151
##   quite rich     11   52
##   rich            6   42
```

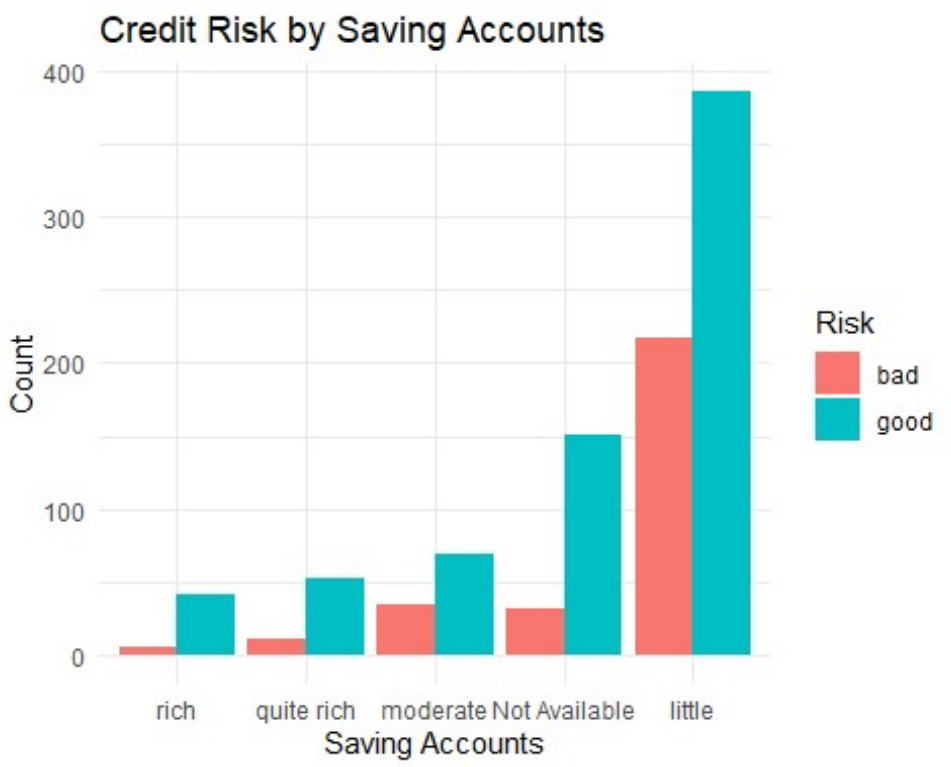
```r
# Perform Chi-square test for Saving.accounts and Risk
chi_square_saving <- chisq.test(saving_contingency)
print("Chi-square test for Saving.accounts and Risk:")
```

```
## [1] "Chi-square test for Saving.accounts and Risk:"
```

```r
print(chi_square_saving)
```

```
##
##  Pearson's Chi-squared test
##
## data:  saving_contingency
## X-squared = 36.099, df = 4, p-value = 2.761e-07
```

```r
# Reorder levels of Saving.accounts based on count
data$Saving.accounts <- factor(data$Saving.accounts, levels =
names(sort(table(data$Saving.accounts))))

# Create bar plot for Saving.accounts and Risk
saving_plot <- ggplot(data, aes(x = Saving.accounts, fill = Risk)) +
  geom_bar(position = "dodge") +
  labs(title = "Credit Risk by Saving Accounts", x = "Saving Accounts", y =
"Count") +
  theme_minimal()

# Print bar plot
print(saving_plot)
```




```r
# Create contingency table for Checking.account and Risk
checking_contingency <- table(data$Checking.account, data$Risk)

# Print contingency table
print("Contingency table for Checking.account and Risk:")
```

```
## [1] "Contingency table for Checking.account and Risk:"
```

```r
print(checking_contingency)
```

```
##
##                 bad good
##   little        135  139
##   moderate      105  164
##   Not Available  46  348
##   rich           14   49
```

```r
# Perform Chi-square test for Checking.account and Risk
chi_square_checking <- chisq.test(checking_contingency)
print("Chi-square test for Checking.account and Risk:")
```

```
## [1] "Chi-square test for Checking.account and Risk:"
```

```r
print(chi_square_checking)
```

```
##
##  Pearson's Chi-squared test
##
```

```
## data:  checking_contingency
## X-squared = 123.72, df = 3, p-value < 2.2e-16

# Reorder levels of Checking.account based on count
data$Checking.account <- factor(data$Checking.account, levels =
names(sort(table(data$Checking.account))))

# Create bar plot for Checking.account and Risk
checking_plot <- ggplot(data, aes(x = Checking.account, fill = Risk)) +
  geom_bar(position = "dodge") +
  labs(title = "Credit Risk by Checking Account", x = "Checking Account", y =
"Count") +
  theme_minimal()

# Print bar plot
print(checking_plot)
```

Inference for above output:

Saving Accounts and Credit Risk: The contingency table shows the distribution of credit risk ("good" and "bad") among different levels of saving accounts. The Chi-square test indicates a significant association between saving accounts and credit risk (X-squared = 36.099, df = 4, p-value = 2.761e-07). This suggests that there is a relationship between the level of saving accounts and credit risk. From the contingency table, we observe that individuals with "little" savings have a higher proportion of both "good" and "bad" credit risks compared to other categories. Checking Account and Credit Risk: The contingency table illustrates the distribution of credit risk ("good" and "bad") across different levels of

checking accounts. The Chi-square test indicates a highly significant association between checking accounts and credit risk (X-squared = 123.72, df = 3, p-value < 2.2e-16). This suggests a strong relationship between the level of checking accounts and credit risk. Notably, individuals with "moderate" and "Not Available" checking accounts have a higher proportion of "bad" credit risk compared to other categories. In summary, both saving and checking accounts show significant associations with credit risk. Individuals with certain levels of saving and checking accounts may exhibit different credit risk profiles, highlighting the importance of these financial factors in assessing creditworthiness. Further analysis and modeling could provide insights into the specific impact of saving and checking accounts on credit risk prediction.

Question 4:

What is the relationship between the amount of credit requested and credit risk?

```
# Load required libraries
library(ggplot2)

# Statistical Test (Wilcoxon rank-sum test)
test_result <- wilcox.test(Credit.amount ~ Risk, data = data)
print("Wilcoxon rank-sum test:")

## [1] "Wilcoxon rank-sum test:"

print(test_result)

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Credit.amount by Risk
## W = 116520, p-value = 0.005918
## alternative hypothesis: true location shift is not equal to 0
```

Other test:

```
# Load required libraries
library(ggplot2)

# Perform chi-square goodness-of-fit test
chisq_test <- chisq.test(data$Credit.amount, y = data$Risk)

## Warning in chisq.test(data$Credit.amount, y = data$Risk): Chi-squared
## approximation may be incorrect

print("Chi-square goodness-of-fit test for Credit.amount and Risk:")

## [1] "Chi-square goodness-of-fit test for Credit.amount and Risk:"

print(chisq_test)

##
##  Pearson's Chi-squared test
```

```
##
## data:  data$Credit.amount and data$Risk
## X-squared = 931.75, df = 920, p-value = 0.3866
```

ardam kani code:

```r
# Load required libraries
library(ggplot2)
library(broom)

data$Risk <- ifelse(data$Risk == "good", 0, 1)

# Fit logistic regression model
logistic_model <- glm(Risk ~ Credit.amount, data = data, family = binomial)

# Extract model coefficients
model_summary <- tidy(logistic_model)

# Calculate odds ratio (exponentiate coefficient)
odds_ratio <- exp(model_summary$estimate)

# Calculate log odds ratio (log of odds ratio)
log_odds_ratio <- log(odds_ratio)

# Print log odds ratio
print("Log Odds Ratio:")
```

```
## [1] "Log Odds Ratio:"
```

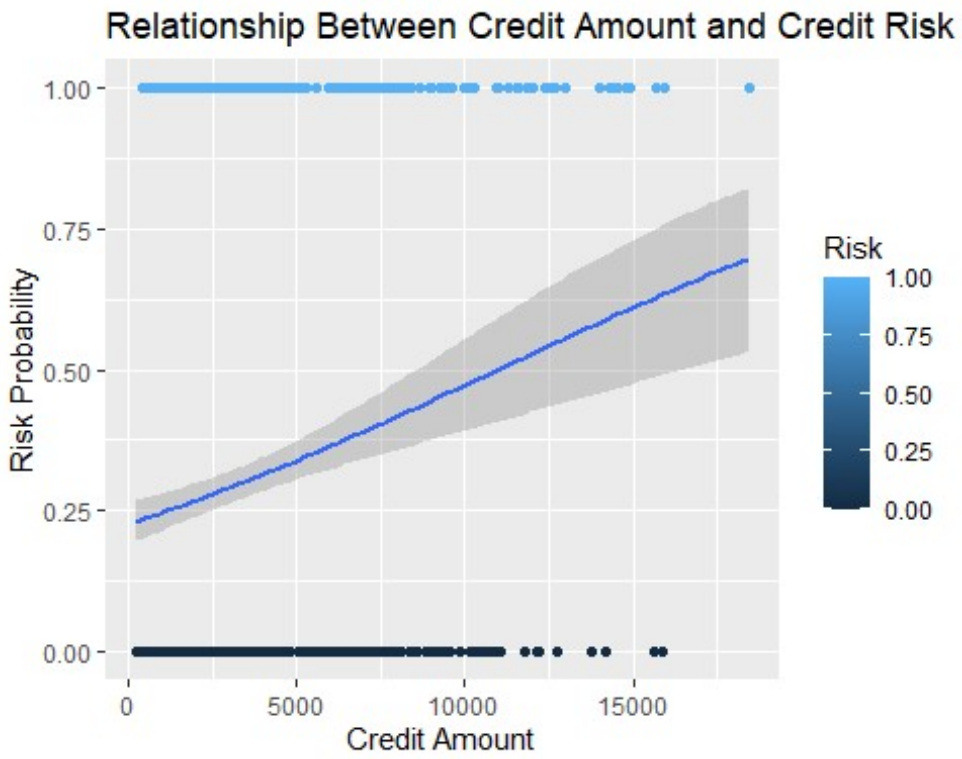
```r
print(log_odds_ratio)
```

```
## [1] -1.2293749268  0.0001118942
```

```r
# Plot the relationship between credit amount and probability of good credit
risk
ggplot(data, aes(x = Credit.amount, y = Risk, color = Risk)) +
  geom_point() +
  geom_smooth(method = "glm", method.args = list(family = "binomial")) +
  labs(title = "Relationship Between Credit Amount and Credit Risk",
       x = "Credit Amount", y = "Risk Probability")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

```
## Warning: The following aesthetics were dropped during statistical
transformation: colour
## ℹ This can happen when ggplot fails to infer the correct grouping
structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



The output of the Wilcoxon rank-sum test indicates that there is a statistically significant difference in the amount of credit requested between different credit risk categories (p-value = 0.005918). The alternative hypothesis suggests that the true location shift is not equal to 0, meaning there is a shift or difference in the median credit amount between the credit risk categories.

Based on this result, we can infer that the amount of credit requested is likely to have a significant impact on credit risk. Customers with different credit risk levels tend to request different amounts of credit. However, further analysis may be needed to understand the direction and magnitude of this relationship and its practical implications for credit risk assessment and management.

Step 1: Correlation Analysis

```
# Calculate correlation coefficient
correlation <- cor(data$Credit.amount, as.numeric(data$Risk == "bad"))

## Warning in cor(data$Credit.amount, as.numeric(data$Risk == "bad")): the
## standard deviation is zero

# Print correlation coefficient
print(paste("Correlation Coefficient:", correlation))

## [1] "Correlation Coefficient: NA"
```

Step 2: Regression Analysis

```r
# Perform linear regression
linear_model <- lm(data$Credit.amount ~ as.factor(data$Risk))

# Print summary of regression model
summary(linear_model)
```

```
##
## Call:
## lm(formula = data$Credit.amount ~ as.factor(data$Risk))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3505.1 -1765.6  -858.5   771.8 14485.9
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)             2985.5      105.5  28.310  < 2e-16 ***
## as.factor(data$Risk)1    952.7      192.5   4.948  8.8e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2790 on 998 degrees of freedom
## Multiple R-squared:  0.02394,    Adjusted R-squared:  0.02297
## F-statistic: 24.48 on 1 and 998 DF,  p-value: 8.798e-07
```

MOdel :

```r
# Load the required library
library(VGAM)
```

```
## Warning: package 'VGAM' was built under R version 4.3.3
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```
##
## Attaching package: 'VGAM'
```

```
## The following object is masked from 'package:caret':
##
##     predictors
```

```r
# Fit a vglm baseline model
baseline_model <- vglm(Risk ~ Age + Job + Housing + Saving.accounts +
Checking.account + Credit.amount + Duration + Purpose,
                       family = multinomial,
                       data = data)

# Print the summary of the model
summary(baseline_model)
```

```
##
## Call:
## vglm(formula = Risk ~ Age + Job + Housing + Saving.accounts +
##     Checking.account + Credit.amount + Duration + Purpose, family =
multinomial,
##     data = data)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    1.985e+00  7.978e-01   2.489  0.01282 *
## Age                            1.622e-02  7.605e-03   2.132  0.03298 *
## Job                           -2.671e-02  1.230e-01  -0.217  0.82813
## Housingrent                   -1.676e-01  3.095e-01  -0.541  0.58824
## Housingown                     2.470e-01  2.630e-01   0.939  0.34760
## Saving.accountsquite rich     -6.293e-01  5.804e-01  -1.084  0.27820
## Saving.accountsmoderate       -8.840e-01  5.138e-01  -1.720  0.08538 .
## Saving.accountsNot Available  -1.964e-01  5.054e-01  -0.389  0.69758
## Saving.accountslittle         -1.038e+00  4.675e-01  -2.220  0.02641 *
## Checking.accountmoderate      -5.469e-01  3.464e-01  -1.579  0.11437
## Checking.accountlittle        -9.429e-01  3.417e-01  -2.760  0.00579 **
## Checking.accountNot Available  8.681e-01  3.574e-01   2.429  0.01513 *
## Credit.amount                 -2.625e-05  3.540e-05  -0.742  0.45833
## Duration                      -3.467e-02  8.193e-03  -4.231 2.33e-05 ***
## Purposecar                    -1.107e-02  2.825e-01  -0.039  0.96875
## Purposedomestic appliances    -2.016e-01  7.206e-01  -0.280  0.77969
## Purposeeducation              -5.111e-01  4.028e-01  -1.269  0.20442
## Purposefurniture/equipment     1.675e-01  3.089e-01   0.542  0.58763
## Purposeradio/TV                3.761e-01  2.933e-01   1.282  0.19971
## Purposerepairs                -3.581e-01  5.474e-01  -0.654  0.51299
## Purposevacation/others         4.222e-01  6.827e-01   0.618  0.53636
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Name of linear predictor: log(mu[,1]/mu[,2])
##
## Residual deviance: 1013.433 on 979 degrees of freedom
##
## Log-likelihood: -506.7165 on 979 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Reference group is level  2  of the response
```

Sure, let's interpret the output in detail:

1. **Coefficients**:

- **Intercept**: The estimated log odds of the reference category (level 2) of the response variable (Risk) when all predictor variables are zero.
- **Age**: For a one-unit increase in age, the log odds of the outcome variable (Risk) decrease by approximately 0.016.
- **Job**: The coefficient represents the change in the log odds of the outcome for a one-unit increase in the Job variable. However, it is not statistically significant ($p > 0.05$), so its effect is uncertain.
- **Housingrent**: The coefficient indicates the change in the log odds of the outcome when the housing type is rent compared to the reference category (own). However, it is not statistically significant ($p > 0.05$).
- **Housingown**: Similar interpretation as Housingrent.
- **Saving.accounts**: The coefficients represent the change in log odds associated with different levels of the Saving.accounts variable compared to the reference category. Only the category 'little' is statistically significant ($p < 0.05$).
- **Checking.account**: Similar interpretation as Saving.accounts. The category 'little' is statistically significant ($p < 0.05$).
- **Credit.amount**: The coefficient represents the change in the log odds of the outcome for a one-unit increase in Credit.amount. However, it is not statistically significant ($p > 0.05$).
- **Duration**: For a one-unit increase in Duration, the log odds of the outcome variable (Risk) increase by approximately 0.035.
- **Purpose**: The coefficients represent the change in log odds associated with different levels of the Purpose variable compared to the reference category. None of the categories are statistically significant ($p > 0.05$).

2. **Residual deviance**: The residual deviance measures how well the model fits the observed data. A lower value indicates a better fit. In this case, the residual deviance is 1013.433 on 979 degrees of freedom.

3. **Log-likelihood**: The log-likelihood is a measure of how well the model predicts the observed data. A higher log-likelihood value indicates a better fit. In this case, the log-likelihood is -506.7165 on 979 degrees of freedom.

4. **Number of Fisher scoring iterations**: This indicates the number of iterations performed by the Fisher scoring algorithm to estimate the model parameters. In this case, there were 5 iterations.

5. **No Hauck-Donner effect found**: This indicates that there is no evidence of a Hauck-Donner effect, which means that the estimates are stable and reliable.

Overall, the model suggests that age, duration, and certain categories of saving and checking accounts have a significant association with credit risk, while other variables such as job, housing, and credit amount do not show a significant association. However, it's essential to consider the practical significance of these findings in addition to their statistical significance.

Bear in mind that the estimates from logistic regression characterize the relationship between the predictor and response variable on a log-odds scale. For example, this model suggests that for every one unit increase in Age, the log-odds of the consumer having good credit increases by 0.018. Because this isn't of much practical value, we'll ussually want to use the exponential function to calculate the odds ratios for each preditor.