498818_SDS496_HW1

2024-09-05

HW1 SDS 496: Topics in Statistics: Machine Learning Methods

Problem 1

i. The sample size n is extremely large, and the number of predictors p is small.

We want to use a **flexible statistical learning method** in this case. When the sample size n is extremely large, the risk of overfitting decreases because the model has more data to learn from. Flexible methods tend to perform better in this scenario because there is enough data to balance the complexity of more parameters. Moreover, the Central Limit Theorem suggests that **with a large sample size, the distribution of parameter estimates becomes more stable, further reducing the risk of overfitting.** Additionally, with a small number of predictors, flexible methods won't have too many parameters to estimate, reducing the risk of overfitting while allowing the method to capture more complex relationships within the data than inflexible methods.

ii. The number of predictors p is extremely large, and the number of observations n is small.

When the number of predictors p is extremely large, and the number of observations n is small, it is better to use an **inflexible statistical learning method.** Flexible methods are prone to **overfitting** in this scenario because they have many parameters to estimate, which allows them to capture not only the underlying relationships but also the noise in the data. While flexible methods might fit the training data well, their predictive power on new data is likely to be poor due to overfitting. With limited observations, flexible methods struggle to generalize, leading to a higher risk of fitting irrelevant patterns. In contrast, **inflexible methods**, which impose more structure and involve fewer parameters, help reduce overfitting by limiting the methods's complexity. These methods are less likely to capture noise, resulting in more stable and generalizable predictions, particularly when the number of predictors exceeds the number of observations. In summary, **inflexible methods are preferable when p > n**, where p is the number of predictors and n is the number of observations, because **they reduce complexity and the risk of overfitting**.

iii. The relationship between the predictors and response is highly non-linear.

In this scenario, we would generally expect a **flexible statistical learning method** to perform better because flexible methods can capture non-linear relationships more effectively. Inflexible methods, like linear regression, are limited to simple, linear relationships, while **flexible methods can adapt to the non-linearity in the data, reducing bias and improving predictive performance.**

iv. The variance of the error terms, i.e. Var(ϵ) = σ^2, is extremely high.

When the variance of the error terms is extremely high, an **inflexible statistical learning method** is generally preferable. High variance of the error terms means that there is a lot of noise in the data, which makes it harder for the method to determine the underlying relationship between the predictor and the response. Flexible methods, which can fit more complex patterns, are more likely to capture this noise, leading to overfitting. **Inflexible methods impose stronger assumptions and a simpler structure on the data, which helps them avoid overfitting to the noise.** This makes inflexible methods more robust in the presence of high variance in the error terms, leading to more stable and reliable predictions. In summary, when the error variance is extremely high, inflexible models are preferable because they are less likely to overfit and are better suited for generalizing in the presence of noisy data.

Problem 2

i. We are interested in the photosynthesis data set discussed in class. For each plant, we have measurements on its warming treatment, site location, soil water content and photosynthesis level. We are interested in understanding which factors affect photosynthesis level.

This is a **regression problem** because the goal is to understand how the predictors, (warming treatment, site location, and soil water content) affect the photosynthesis level, which is likely a continous numerical variable.

We are most interested in **inference** because the focus is on understanding which factors affect the photosynthesis level, rather than just predicting photosynthesis levels for new plants. Inference involves understanding the relationships between predictors and the response.

Finally, provide the sample size n and the number of predictors p.

```r
photo = read.csv("/Users/aidanashrafi/Downloads/Reich2018NaturePaperDataAug2018.csv")
sample_size_photo<- nrow(photo)
sample_size_photo
```

```
## [1] 2052
```

p = 3, n = 2052

ii. We have synthesized a new drug, and we wish to know whether it will be successful or not at treating illness X. We collect data on 20 similar drugs designed to treat illness X. For each drug we have recorded whether it was successful or not and a list of its active ingredients.

This is a **classification problem** because the response variable is categorical. The goal is to predict whether a drug will be successful or not at treating illness which is a categorical outcome (binary classification).

We are most interested in **prediction** because the focus is on predicting whether the new drug will be successful or not, based on data from similar drugs. This focuses on using the model to predict outcomes for future instances, rather than understanding the relationship between each active ingredient and success.

n = 20, p = distinct number of active ingredients across the 20 drugs

Problem 3

i. What are the advantages and disadvantages of a very flexible approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

Advantages of a very flexible approach: captures complex relationships, lower bias. Flexible models can capture non-linear patterns in the data that less flexible models might miss. They generally have lower bias because they fit the data more closely.

Disadvantages of a very flexible approach: overfitting, requires large number of data points, and can be computationally slow/complex. Flexible models tend to overfit especially when the data set is small or noisy. Overfitting means the model additionally captures the noise or random fluctuations in the data which leads to poor insights on new, unseen data.

When might a more flexible approach be preferred? When the relationship is non-linear or complex, and when working with large data sets, as large data sets make flexible models less likely to overfit.

When might a less flexible approach be preferred? When there are small sample sizes, and simple relationships, inflexible methods like linear regression may be preferred. Linear regression allows us to generalize better with limited data because it reduces the risk of overfitting with data that has simple/approximately linear relationships.

ii. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification? What are its disadvantages?

Parametric statistical learning approach is one that maps a previously known function to the data, and takes on certain assumptions and functional forms. Intuitively, one can see the potential statistical power in predicting with a parametric approach if functional form assumptions are able to hold and the function has a high goodness of fit to the data.

Non-parametric statistical learning approach is one that does not assume a fixed functional form. Instead, they rely on the data itself to estimate relationships between predictors and response, making them more flexible and adaptable. They allow the model to be more flexible, and data-driven in estimating the relationship between predictors and response.

Advantages of Parametric: Simpler and more interpretable, less data required, and computationally efficient. Parametric models are often easier to understand and interpret, require fewer data points to estimate parameters, and are computationally efficient to fit.

Disavantages of Parametric: Assumes a functional form, limited flexibility. If the assumed functional form of the model is incorrect, the model will suffer from high bias and may underfit. Parametric models are not well-suited for capturing complex, non-linear relationships.

Advantages of Non-Parametric: Very flexible, predictive insights. Non-parametric methods can model complex, and non-linear relationships without making strong assumptions about the data. They can generate predictions in cases where parametric models fail to fit the data well, because non-parametric methods do not assume a specific functional form.

Disadvantages of Non-Parametric: High Variance and risk of overfitting, requires large data sets, interpretability. Since Non-Parametric methods are flexible enough to capture noise in the data, they are prone to overfitting, especially with small data sets. Non-parametric methods need more data to perform well and to avoid overfitting. Non-Parametric methods interpretability can be lower compared to parametric models, especially in complex methods.

Problem 4

```r
#install.packages('tidyverse')
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.0     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```r
# Load the dataset in personal console
# WARNING: Need to change file path if you want to run on your computer
data = read.csv("/Users/aidanashrafi/Downloads/horse-colic.csv")
# take a quick look at the dataset
head(data)
```

```
##   surgery Age Hospital.Number rectal.temperature pulse respiratory.rate
## 1       2   1          530101               38.5    66               28
## 2       1   1          534817               39.2    88               20
## 3       2   1          530334               38.3    40               24
## 4       1   9         5290409               39.1   164               84
## 5       2   1          530255               37.3   104               35
## 6       2   1          528355                 NA    NA               NA
##   temperature.of.extremities peripheral.pulse mucous.membranes
## 1                          3                3               NA
## 2                         NA               NA                4
## 3                          1                1                3
## 4                          4                1                6
## 5                         NA               NA                6
## 6                          2                1                3
##   capillary.refill.time pain peristalsis abdominal.distension nasogastric.tube
## 1                     2    5           4                    4               NA
## 2                     1    3           4                    2               NA
## 3                     1    3           3                    1               NA
## 4                     2    2           4                    4                1
## 5                     2   NA          NA                   NA               NA
## 6                     1    2           3                    2                2
##   nasogastric.reflux nasogastric.reflux.PH rectal.examination abdomen
## 1                 NA                    NA                  3       5
## 2                 NA                    NA                  4       2
## 3                 NA                    NA                  1       1
## 4                  2                     5                  3      NA
## 5                 NA                    NA                 NA      NA
## 6                  1                    NA                  3       3
##   packed.cell.volume total.protein abdominocentesis.appearance
## 1                 45           8.4                          NA
## 2                 50          85.0                           2
## 3                 33           6.7                          NA
## 4                 48           7.2                           3
## 5                 74           7.4                          NA
## 6                 NA            NA                          NA
##   abdomcentesis.total.protein outcome surgical.lesion type.of.lesion.1
## 1                          NA       2               2            11300
## 2                         2.0       3               2             2208
## 3                          NA       1               2                0
## 4                         5.3       2               1             2208
## 5                          NA       2               2             4300
## 6                          NA       1               2                0
##   type.of.lesion.2 type.of.lesion.3 cp_data
## 1                0                0       2
## 2                0                0       2
## 3                0                0       1
## 4                0                0       1
## 5                0                0       2
## 6                0                0       2
```

```r
# Target the variables of interest, extract all rows of the specific columns
clean_data <- data[, c("respiratory.rate","pulse","rectal.temperature","pain","packed.cell.volume","tota

# Check to make sure operation was performed correctly
head(clean_data)
```

```
##   respiratory.rate pulse rectal.temperature pain packed.cell.volume
## 1               28    66               38.5    5                 45
## 2               20    88               39.2    3                 50
## 3               24    40               38.3    3                 33
## 4               84   164               39.1    2                 48
## 5               35   104               37.3   NA                 74
## 6               NA    NA                 NA    2                 NA
##   total.protein Age abdominal.distension
## 1           8.4   1                    4
## 2          85.0   1                    2
## 3           6.7   1                    1
## 4           7.2   9                    4
## 5           7.4   1                   NA
## 6            NA   1                    2
```

```
initial_sample_size<- nrow(clean_data)
initial_sample_size
```

```
## [1] 300
```

```
# Drop NA values, rename data, two ways to do so
# This will reorder the indexes chronologically after removing NA
cleaner_data<- drop_na(clean_data)
head(cleaner_data)
```

```
##   respiratory.rate pulse rectal.temperature pain packed.cell.volume
## 1               28    66               38.5    5                 45
## 2               20    88               39.2    3                 50
## 3               24    40               38.3    3                 33
## 4               84   164               39.1    2                 48
## 5               16    48               37.9    3                 37
## 6               12    66               38.1    3                 44
##   total.protein Age abdominal.distension
## 1           8.4   1                    4
## 2          85.0   1                    2
## 3           6.7   1                    1
## 4           7.2   9                    4
## 5           7.0   1                    3
## 6           6.0   1                    1
```

```
cleaner_sample_size <- nrow(cleaner_data)
cleaner_sample_size
```

```
## [1] 147
```

```
# This will leave all indexes in its original position
cleaned_data <- na.omit(clean_data)
head(cleaned_data)
```

```
##    respiratory.rate pulse rectal.temperature pain packed.cell.volume
## 1                28    66               38.5    5                 45
## 2                20    88               39.2    3                 50
## 3                24    40               38.3    3                 33
## 4                84   164               39.1    2                 48
## 7                16    48               37.9    3                 37
## 11               12    66               38.1    3                 44
##    total.protein Age abdominal.distension
## 1            8.4   1                    4
## 2           85.0   1                    2
## 3            6.7   1                    1
## 4            7.2   9                    4
## 7            7.0   1                    3
## 11           6.0   1                    1
```

```
cleaned_sample_size<-nrow(cleaned_data)
cleaned_sample_size
```

```
## [1] 147
```

i. Read the data set and inspect the variables of interest. What's the sample size? What's the remaining sample size after taking out every data entry with a missing value in at least one of the variables of interest?

The sample size of the data set considering only the variables of interest **(respiratory.rate, pulse, rectal.temperature, pain, packed.cell.volume, total.protein, Age, abdominal.distension) was 300**. The remaining sample size after taking out every data entry with a missing value in at least one of the variables of interest was **147**.

ii. Which of the predictors are quantitative and which are qualitative?

```
str(cleaned_data)
```

```
## 'data.frame':    147 obs. of  8 variables:
##  $ respiratory.rate    : int  28 20 24 84 16 12 52 28 28 48 ...
##  $ pulse               : int  66 88 40 164 48 66 72 92 76 96 ...
##  $ rectal.temperature  : num  38.5 39.2 38.3 39.1 37.9 38.1 39.1 38 38.2 37.6 ...
##  $ pain                : int  5 3 3 2 3 3 2 1 3 5 ...
##  $ packed.cell.volume  : num  45 50 33 48 37 44 50 37 46 45 ...
##  $ total.protein       : num  8.4 85 6.7 7.2 7 6 7.8 6.1 81 6.8 ...
##  $ Age                 : int  1 1 1 9 1 1 1 9 1 1 ...
##  $ abdominal.distension: int  4 2 1 4 3 1 2 2 1 3 ...
##  - attr(*, "na.action")= 'omit' Named int [1:153] 5 6 8 9 10 13 17 18 20 24 ...
##   ..- attr(*, "names")= chr [1:153] "5" "6" "8" "9" ...
```

Quantitative Variables: These variables take on continuous or discrete numerical values such as measurements or counts. In the dataset, **respiratory.rate, pulse, rectal.temperature, packed.cell.volume, and total.protein are quantitative variables** despite some being stored as integers. These variables are quantitative because they measure continuous characteristics (rates,volumes, temperatures).

Qualitatitve Variables: These variables represent categories or labels, even if they are stored as integers. In the dataset, **pain, age, abdominal.distension are qualitative variables**. These variables are qualitative because they have discrete characteristics and represent distinct, categorical values rather than continous measurements.

Create a summary of some standard statistics for each of the quantitative variables. Create a contingency table exploring the distribution of counts among the classes of each qualitative variable.

```
quant_data<- cleaned_data[,c('respiratory.rate','pulse','rectal.temperature','packed.cell.volume','tota
quant_summary<- summary(quant_data)
qual_data<- cleaned_data[,c('pain','Age','abdominal.distension')]
pain_summary<- table(qual_data$pain)
age_summary<- table(qual_data$Age)
abdominal_distension_summary<- table(qual_data$abdominal.distension)
quant_summary
```

```
##  respiratory.rate     pulse         rectal.temperature packed.cell.volume
##  Min.   : 8.00    Min.   : 36.00   Min.   :36.40      Min.   :23.00
##  1st Qu.:18.00    1st Qu.: 48.00   1st Qu.:37.80      1st Qu.:38.00
##  Median :24.00    Median : 60.00   Median :38.10      Median :44.00
##  Mean   :29.26    Mean   : 69.84   Mean   :38.17      Mean   :45.61
##  3rd Qu.:36.00    3rd Qu.: 86.00   3rd Qu.:38.50      3rd Qu.:50.00
##  Max.   :88.00    Max.   :184.00   Max.   :40.80      Max.   :75.00
##  total.protein
##  Min.   : 3.3
##  1st Qu.: 6.5
##  Median : 7.3
##  Mean   :23.3
##  3rd Qu.:55.0
##  Max.   :85.0
```

```
cat("\nPain Summary:\n")
```

```
##
## Pain Summary:
```

```
pain_summary
```

```
##
##  1  2  3  4  5
## 26 40 46 20 15
```

```
cat("\nAge Summary:\n")
```

```
##
## Age Summary:
```

```
age_summary
```

```
##
##   1   9
## 136  11
```

```
cat("\nAbdominal Distension Summary:\n")
```

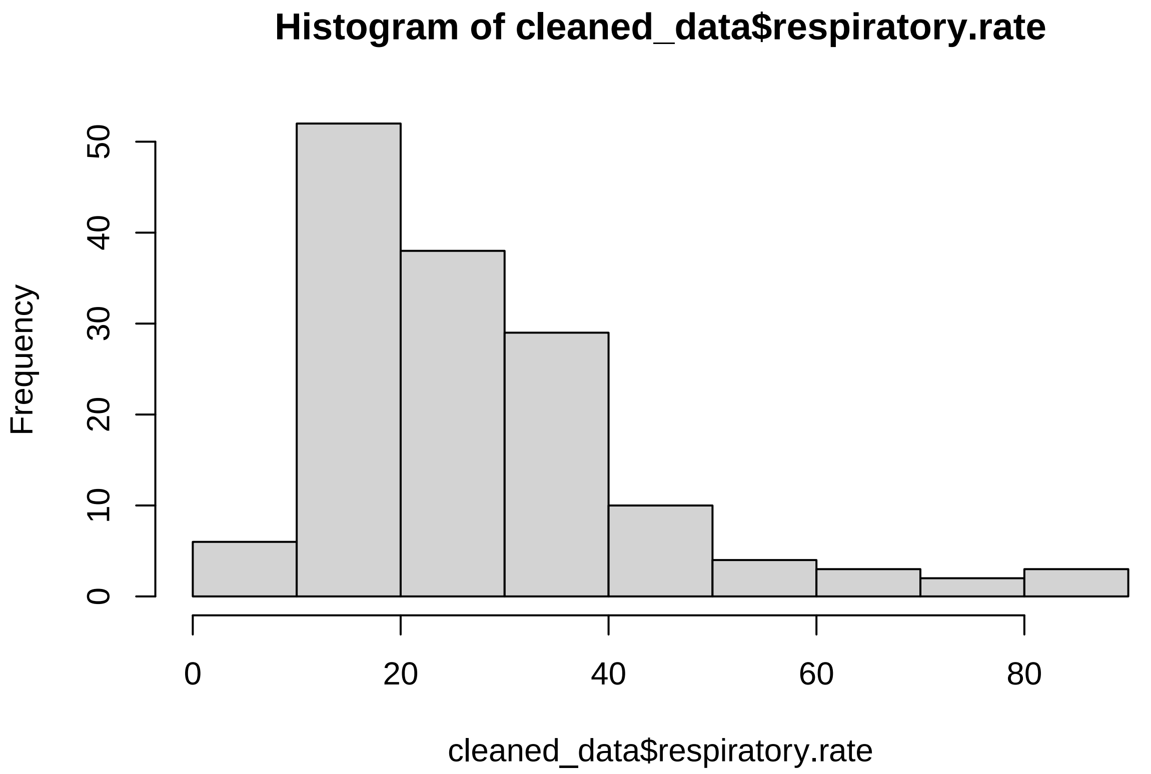
```
##
## Abdominal Distension Summary:
```

```
abdominal_distension_summary
```

```
##
##  1  2  3  4
## 50 42 41 14
```

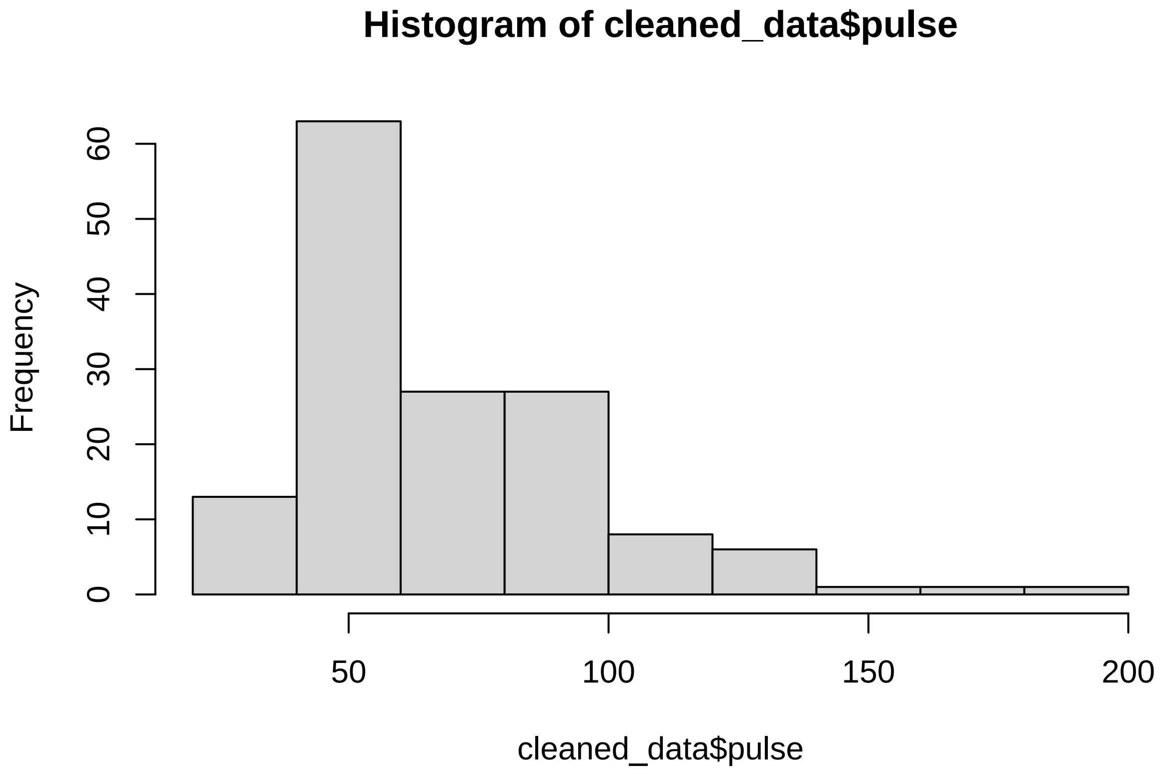
iii. Plot a histogram exploring the distribution of each quantitative variable. Comment on whether each of those distributions appears to be symmetric. If not, check whether a simple transformation of the quantitative variable can lead to a more symmetric distribution.

```
hist(cleaned_data$respiratory.rate)
```



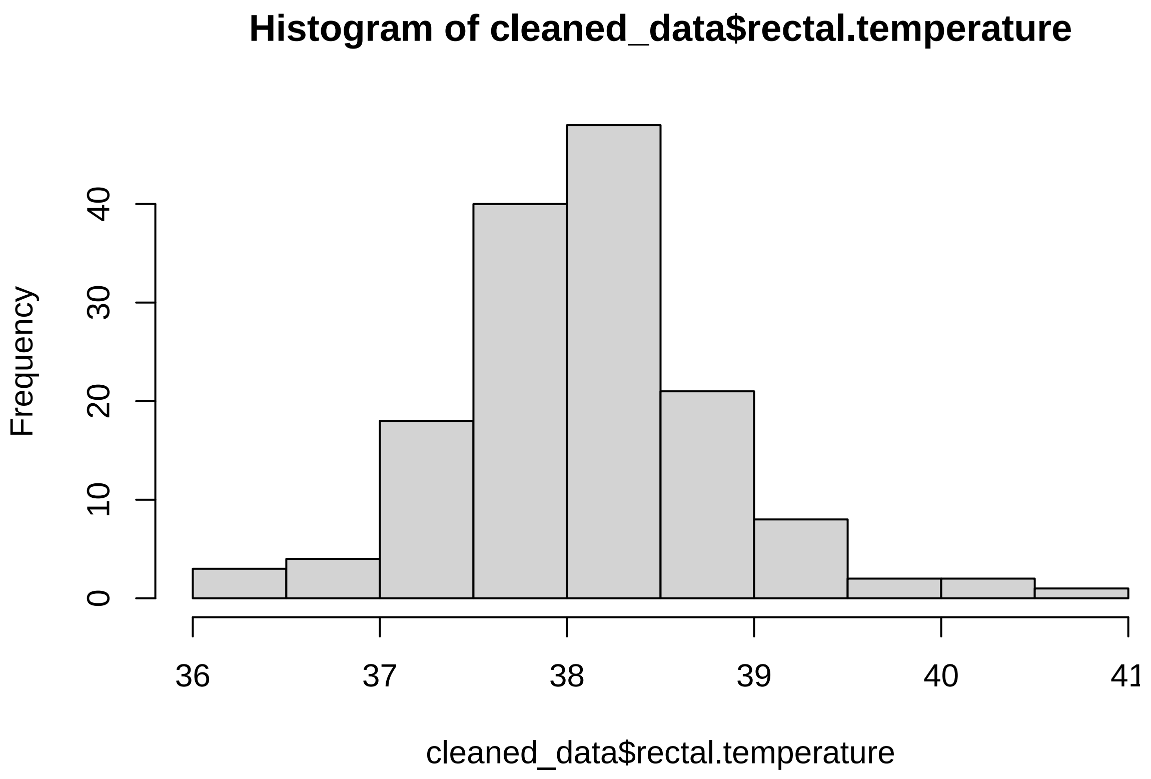
This variable is not symmetric.

```
hist(cleaned_data$pulse)
```



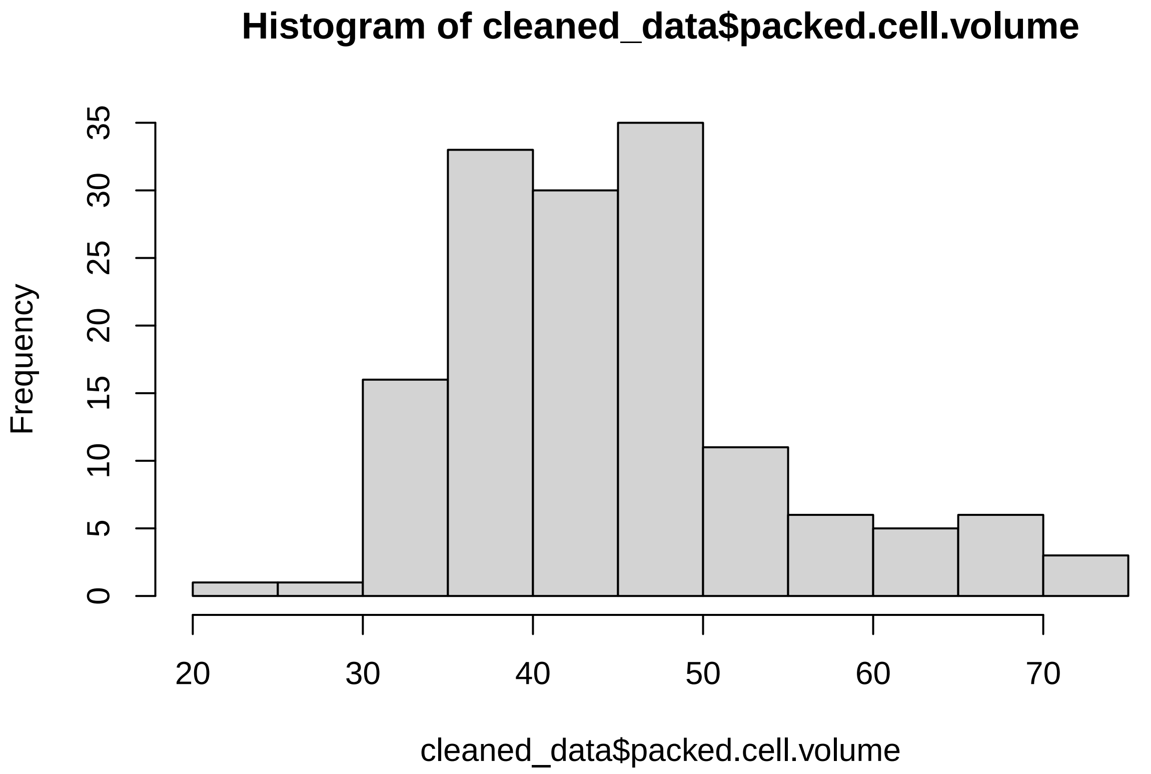
This variable is not symmetric.

```r
hist(cleaned_data$rectal.temperature)
```



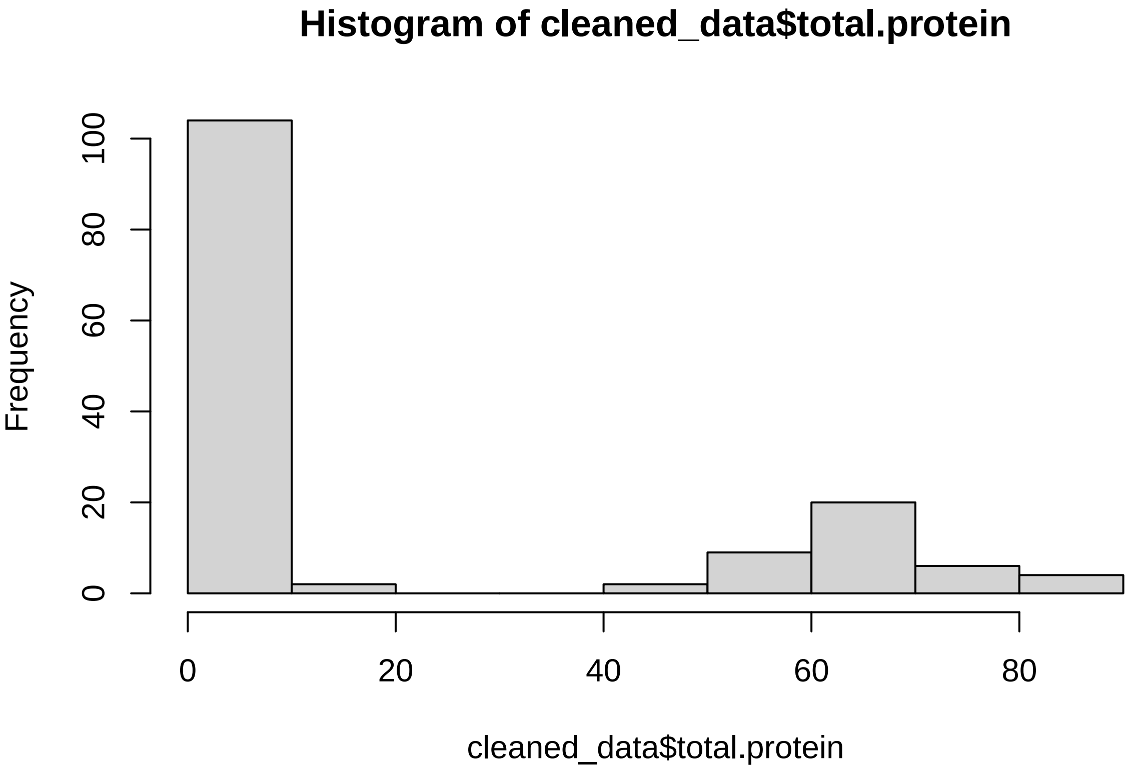
This variable is symmetric.

```r
hist(cleaned_data$packed.cell.volume)
```



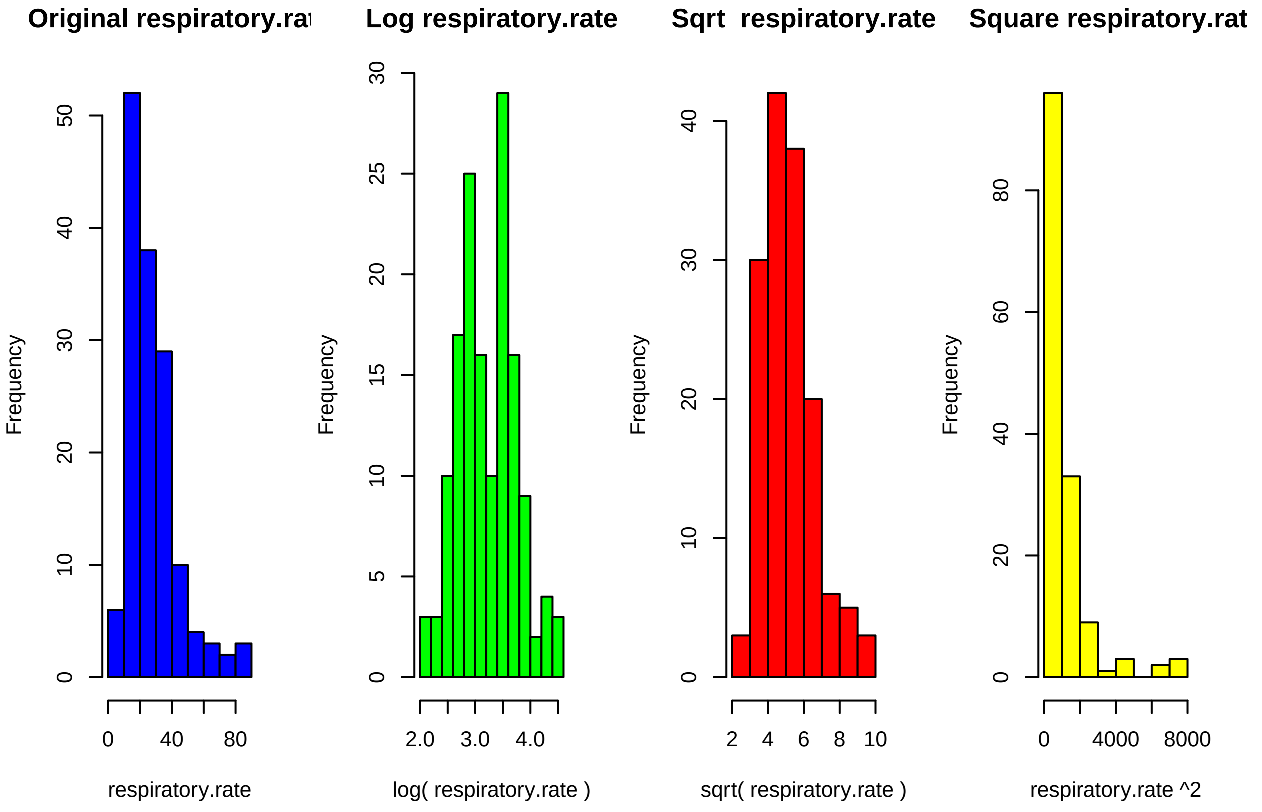
This variable is symmetric.

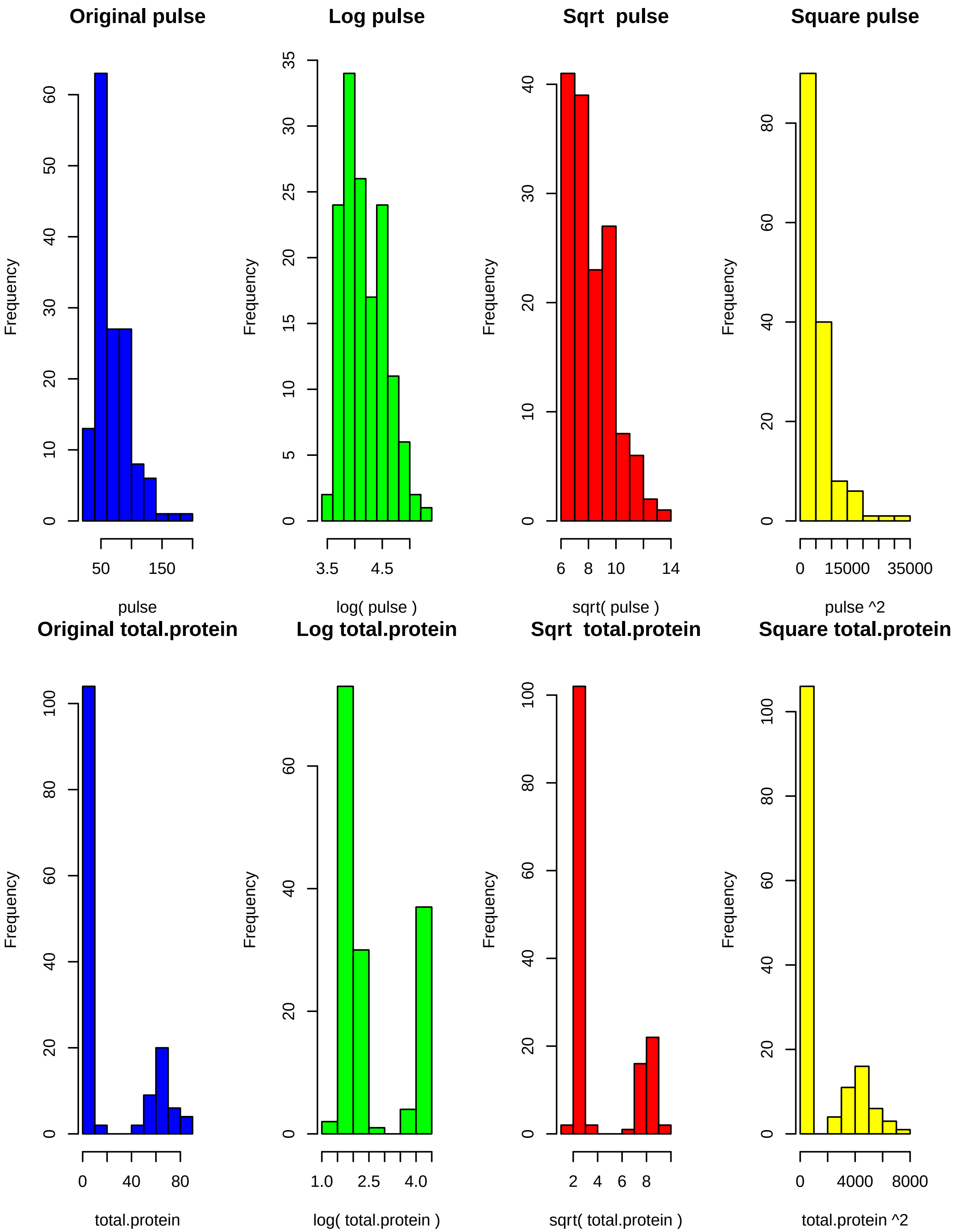
```r
hist(cleaned_data$total.protein)
```



This variable is not symmetric.

```r
# List of asymmetric quantitative variables
asymmetric_quantitative_vars <- c("respiratory.rate", "pulse", "total.protein")

# Loop through each variable and apply log, sqrt, and square transformations
for (i in asymmetric_quantitative_vars) {
  # Log transformation
  log_transformed <- log(cleaned_data[, i])

  # Square root transformation
  sqrt_transformed <- sqrt(cleaned_data[, i])

  # Square transformation
  square_transformed <- cleaned_data[, i]^2

  # Plot the original and transformed data for comparison
  par(mfrow = c(1, 4))  # 1 row, 4 columns

  # Original Histogram
  hist(cleaned_data[, i], main = paste("Original", i), xlab = i, col = "blue")

  # Log-transformed histogram
  hist(log_transformed, main = paste("Log", i), xlab = paste("log(", i,")"), col = "green")

  # Square root-transformed histogram
  hist(sqrt_transformed, main = paste("Sqrt ", i), xlab = paste("sqrt(", i, ")"), col = "red")
  # Square-transformed histogram
  hist(square_transformed, main = paste("Square", i), xlab = paste(i, "^2"), col = "yellow")
}
```




Original pulse
Frequency
0
10
20
30
40
50
60
50
150
pulse
Log pulse
Frequency
0
5
10
15
20
25
30
35
3.5
4.5
log(pulse)
Sqrt pulse
Frequency
0
10
20
30
40
6
8
10
14
sqrt(pulse)
Square pulse
Frequency
0
20
40
60
80
0
15000
35000
pulse ^2
Original total.protein
Frequency
0
20
40
60
80
100
0
40
80
total.protein
Log total.protein
Frequency
0
20
40
60
1.0
2.5
4.0
log(total.protein)
Sqrt total.protein
Frequency
0
20
40
60
80
100
2
4
6
8
sqrt(total.protein)
Square total.protein
Frequency
0
20
40
60
80
100
0
4000
8000
total.protein ^2

Histogram Summary

After applying log, square root, and square transformations to the asymmetric quantitative variables in the dataset, it is clear that the log transformation enables respiratory rate and pulse data to follow a symmetric distribution. Because of the inherent nature of something like total proteins data for medical conditions and the tendency for measurements to be extreme and have a large spread in between extremes, after simple transformations, total.protein was still inherently skewed which makes sense.

iv. Calculate a pairwise correlation coefficient for each combination of 2 (possibly transformed) quantitative variables. Is there any strong linear relationship between any of the predictors and the the response variable? Are there any strong linear relationships among the predictors?

```
# Extract the quantitative variables (possibly transformed) from the dataset
quantitative_vars <- c("respiratory.rate", "pulse", "rectal.temperature", "packed.cell.volume", "total.p

# Calculate the correlation matrix for the quantitative variables
cor_matrix <- cor(cleaned_data[, quantitative_vars], use = "complete.obs")

# Print the correlation matrix
cat("\nPairwise Correlation Matrix:\n")
```

```
##
## Pairwise Correlation Matrix:
```

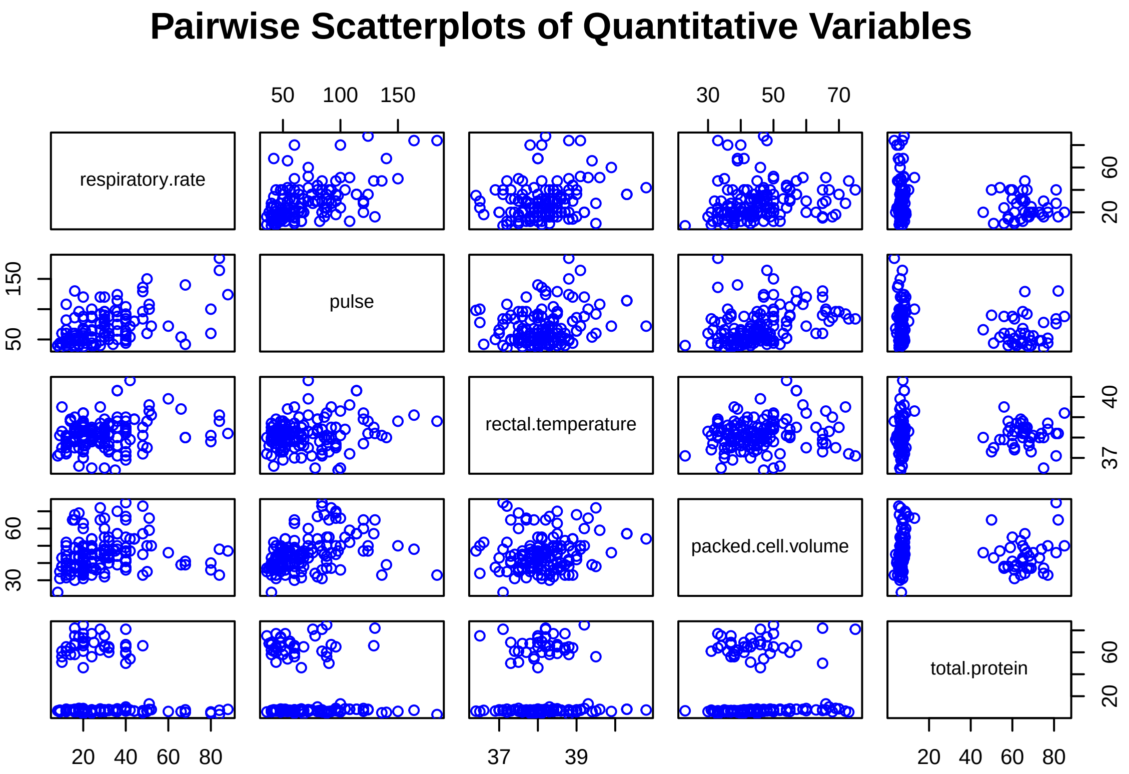
```
print(cor_matrix)
```

```
##                    respiratory.rate      pulse rectal.temperature
## respiratory.rate          1.0000000  0.5391872         0.23424853
## pulse                     0.5391872  1.0000000         0.21803062
## rectal.temperature        0.2342485  0.2180306         1.00000000
## packed.cell.volume        0.1515527  0.3815847         0.16777132
## total.protein            -0.1857121 -0.1468428        -0.01441778
##                    packed.cell.volume total.protein
## respiratory.rate           0.15155272   -0.18571205
## pulse                      0.38158467   -0.14684284
## rectal.temperature         0.16777132   -0.01441778
## packed.cell.volume         1.00000000   -0.05830071
## total.protein             -0.05830071    1.00000000
```

Based on the correlation matrix: **pulse has the strongest positive correlation to respiratory rate** (0.5391872). This shows there is not a particularly powerful/strong linear relationship between any predictor and respiratory rate, all correlations to response are < 0.60. In addition, total.protein is negatively correlated to respiratory rate (-0.18571205), and is the only quantitative predictor negatively correlated with the response. **Packed cell volume and pulse have the highest correlation among predictors** at (0.38158467). There is not a particularly strong relationship among any two predictors either, with the strongest correlation being < 0.40.

v. Create a pairwise scatterplot for each combination of 2 (possibly transformed) quantitative variables. Does any of the predictors appear to be particularly useful in making predictions about the response variable? Do any of the predictors appear to be strongly associated with each other?

```
# Extract the relevant quantitative variables
quantitative_vars <- cleaned_data[, c("respiratory.rate", "pulse", "rectal.temperature", "packed.cell.v
# Create pairwise scatterplots for all combinations of quantitative variables
pairs(quantitative_vars, main = "Pairwise Scatterplots of Quantitative Variables", col = "blue")
```



None of the predictors are strongly, linearly related to the response variable. Our correlation coefficients between the response and each individual predictor were generally low, meaning there are no strong linear correlations between the predictors and the response variable. The scatterplots did not show any strong patterns between the predictors and the response.

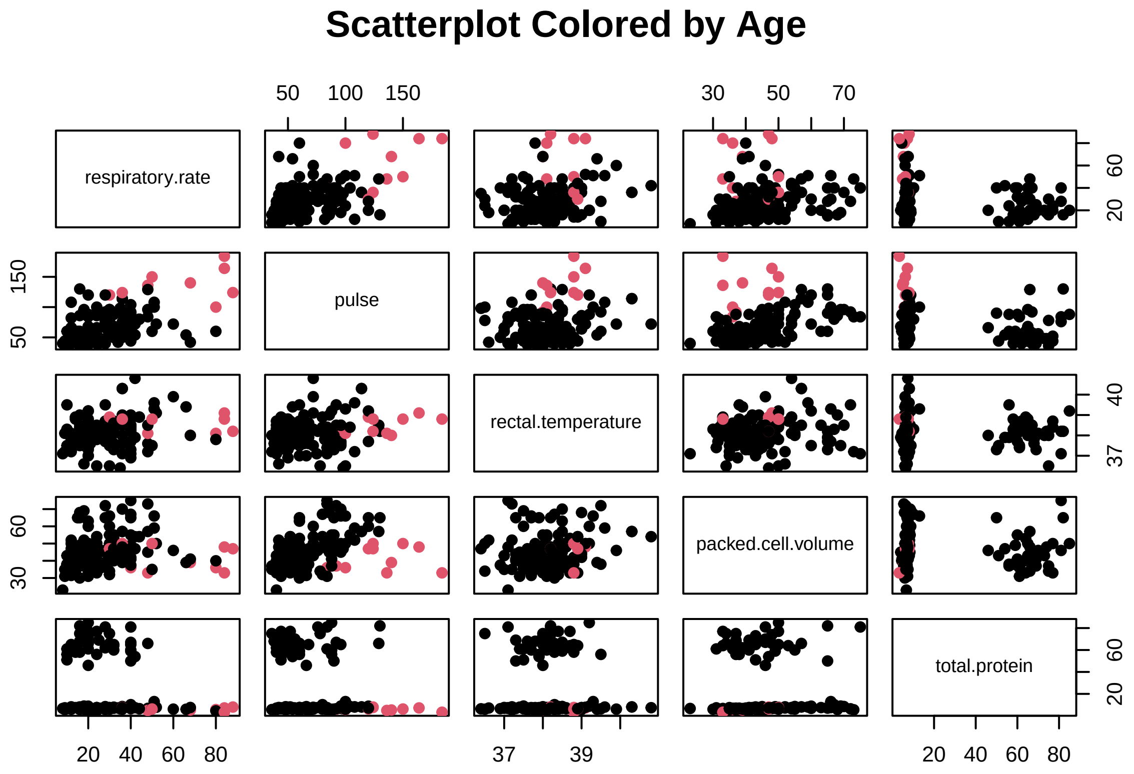
None of the predictors seem to be strongly associated with each other either. Our correlation coefficients across predictors was generally low. The scatterplots did not show any strong patterns among predictors.

vi. Color the pairwise scatterplots according to each of the qualitative variables separately. Are there any obvious clusters of observations corresponding to the classes of some qualitative variable appearing in any of the scatterplots?

Assigns colors to different factor levels in on chronological order based on the levels of the factor.

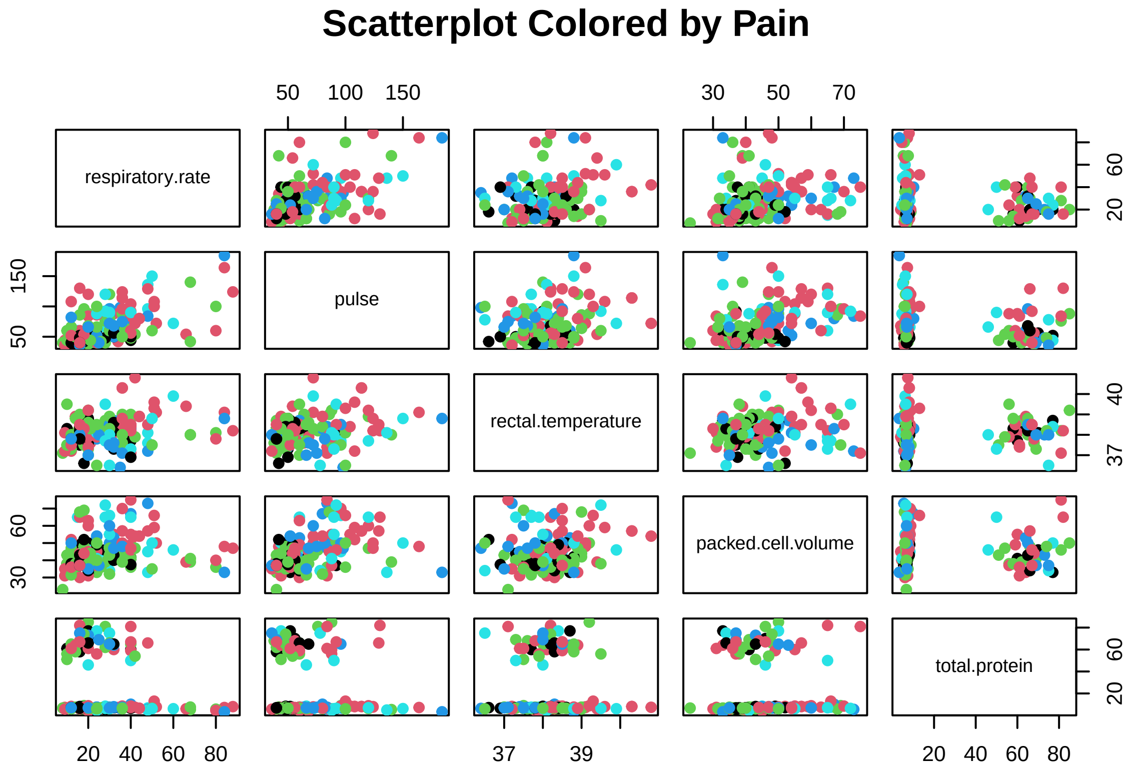
Color Order : Black, Red, Green, Blue, Cyan, Magenta, Yellow, Gray

```
# Assigns colors to different factor levels in on chronological order based on the levels of the factor
# Color Order : Black, Red, Green, Blue, Cyan, Magenta, Yellow, Gray
# 1. Color the scatterplots based on 'Age'
pairs(cleaned_data[, c("respiratory.rate", "pulse", "rectal.temperature", "packed.cell.volume", "total.p
      main = "Scatterplot Colored by Age",
      col = factor(cleaned_data$Age), pch = 19)
```



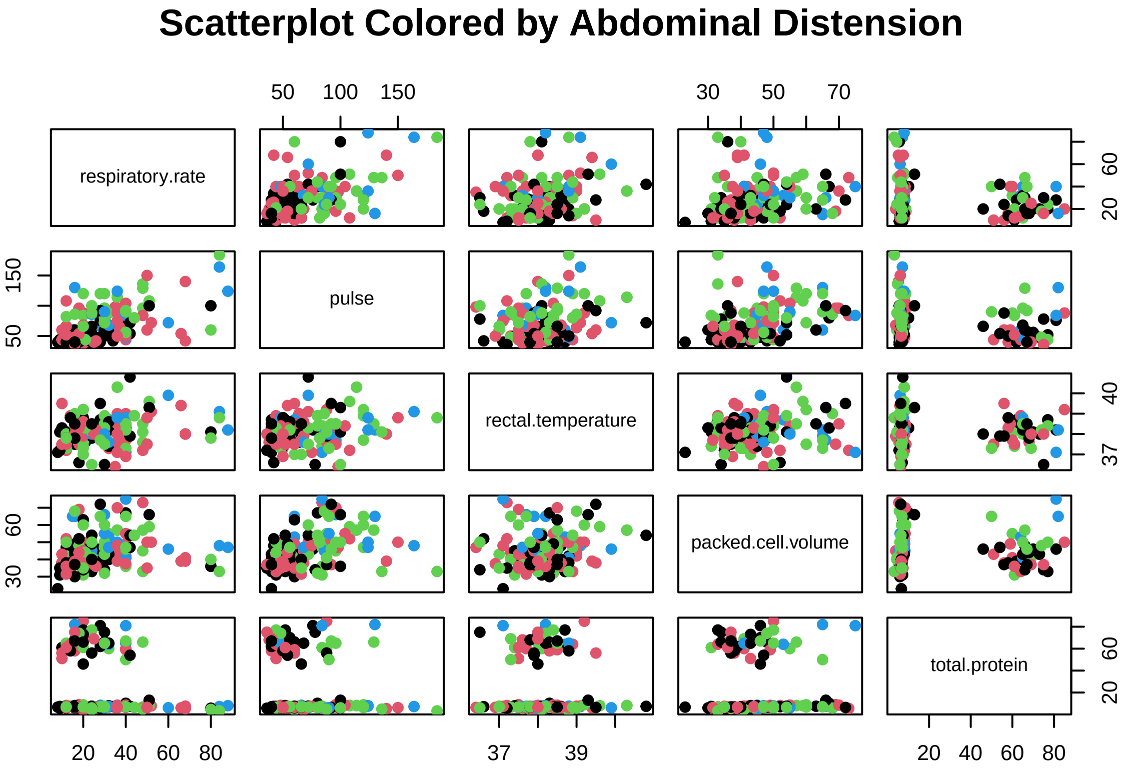
Scatterplot Colored by Age

```r
# 2. Color the scatterplots based on 'Pain'
pairs(cleaned_data[, c("respiratory.rate", "pulse", "rectal.temperature", "packed.cell.volume", "total.
      main = "Scatterplot Colored by Pain",
      col = factor(cleaned_data$pain), pch = 19)
```



Scatterplot Colored by Pain

```r
# 3. Color the scatterplots based on 'Abdominal Distension'
pairs(cleaned_data[, c("respiratory.rate", "pulse", "rectal.temperature", "packed.cell.volume", "total.
      main = "Scatterplot Colored by Abdominal Distension",
      col = factor(cleaned_data$abdominal.distension), pch = 19)
```



The black and red dots are mostly mixed together across the plots, black representing 1, red representing 9, suggesting that Age does not appear to create distinct clusters in relation to the other variables. There is no clear pattern where one age group has distinctly different quantitative measurements than the other.

In the scatterplot colored by Pain, we see that the colors are more evenly distributed across the scatterplots, meaning that the Pain levels (as represented by the different colors) don't show strong separations in the data. Like Age, Pain doesn't seem to explain much of the variability in the quantitative predictors. There is no obvious clustering of colors that would suggest strong differences in the quantitative variables based on pain levels.

In the scatterplot colored by Abdominal Distension, there's a bit more variation in the distribution of colors, but still no clear clusters. The colors (which represent different levels of abdominal distension) are somewhat spread out across the variables. No strong clustering was observed that would indicate a strong association between abdominal distension and the other variables.

New Student, Extension from Professor Note

I was not enrolled in this course for the first 2 weeks so I was granted an extension on the first homework assignment. Furthermore, I was not sure if professor mentioned anything about Problem 2 part ii, in regards to the number of predictors during the first two weeks of class as I was not there, but I answered to the best of my understanding.

If there was a specific code or instruction the professor wanted us to use or perform and it was mentioned during the first two weeks of lecture, please consider that I was not there for it, and worked hard to catch up to other students and complete this as soon as possible.

I was unsure about Question 4 part iv, and v, in terms of for each combination of 2 (possibly transformed), I am was not able to ask clarification before the due-date, but I completed the objectives on all quantitative variables.

Thank you for understanding.