STAT 561: Logistic Regression

Prashanna Raj Pandit | Nazma Vali Shaik | Hema Sai Paruchuri

2025-04-02

1. Absenteesim data

```
absent <- read_excel("Absenteeism_at_work.xls")
names(absent)
```

```
##  [1] "ID"                                "Month_of_absence"
##  [3] "Day_of_the_week"                   "Seasons"
##  [5] "Transportation_expense"            "Distance_from_Residence_to_Work"
##  [7] "Service_time"                      "Age"
##  [9] "Work_load_Average_in_days"         "Education"
## [11] "Son"                               "Pet"
## [13] "Weight"                            "Height"
## [15] "Body_mass_index"                   "Absenteeism_time_in_hours"
```

Recode and Factor Conversion

```
absent <- absent %>%
  mutate(absenteeism = case_when(
    Absenteeism_time_in_hours >= 0 & Absenteeism_time_in_hours <= 20 ~ "Low",
    Absenteeism_time_in_hours > 20 & Absenteeism_time_in_hours <= 40 ~
"Moderate",
    Absenteeism_time_in_hours > 40 ~ "High"
  ))

absent$absenteeism <- factor(absent$absenteeism, levels = c("Low",
"Moderate", "High"))
absent$Day_of_the_week <- factor(absent$Day_of_the_week,
                                 levels = 2:6,
                                 labels = c("Monday", "Tuesday",
"Wednesday", "Thursday", "Friday"))
absent$Seasons <- factor(absent$Seasons,
                         levels = 1:4,
                         labels = c("Summer", "Autumn", "Winter", "Spring"))
absent$Education <- factor(absent$Education,
                           levels = 1:4,
                           labels = c("High School", "Graduate",
"Postgraduate", "Master/Doctor"))

str(absent[c("Month_of_absence", "Day_of_the_week", "Seasons", "Education")])
```

```
## tibble [740 × 4] (S3: tbl_df/tbl/data.frame)
##  $ Month_of_absence: num [1:740] 7 7 7 7 7 7 7 7 7 7 ...
##  $ Day_of_the_week : Factor w/ 5 levels "Monday","Tuesday",..: 2 2 3 4 4 5
```

```
5 5 1 1 ...
##  $ Seasons         : Factor w/ 4 levels "Summer","Autumn",..: 1 1 1 1 1 1
1 1 1 1 ...
##  $ Education       : Factor w/ 4 levels "High School",..: 1 1 1 1 1 1 1 1
1 3 ...
```

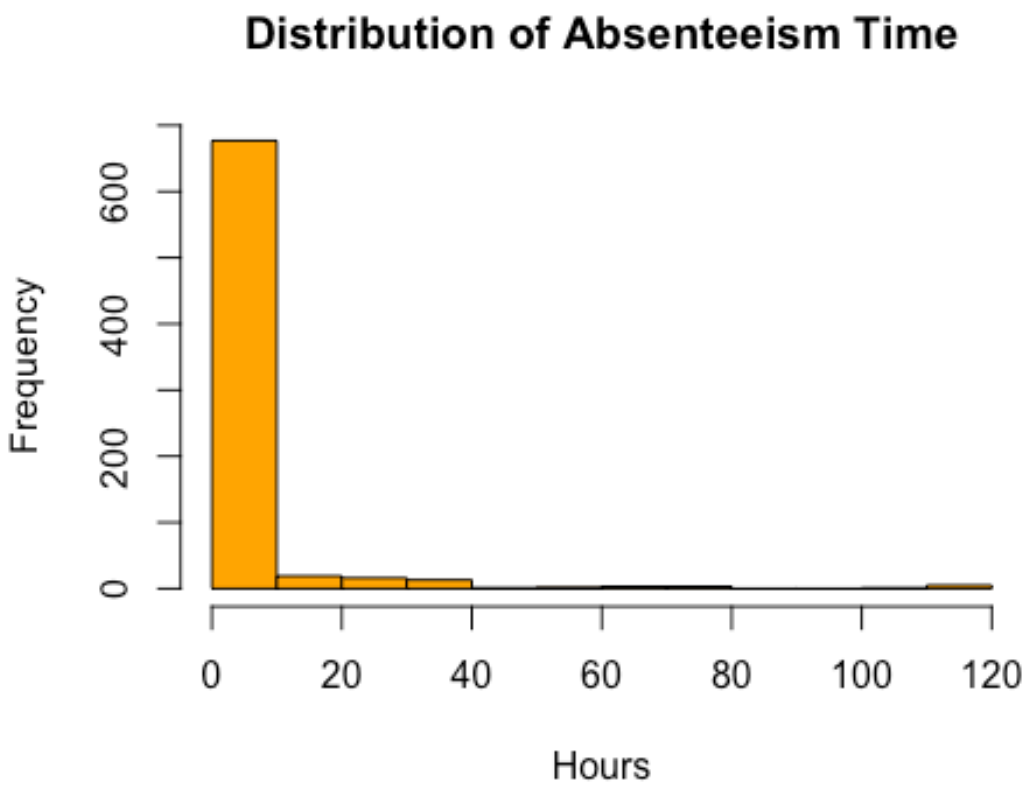
(a) How is the 'Absenteeism time in hours' distributed? Are there any noticeable patterns or outliers?

Interpretation:

In Histogram: The histogram of Absenteeism_time_in_hours reveals a highly right-skewed distribution. Most of the values are clustered at the lower end of the scale. These values suggest that short-term absences (below 10 hours) are extremely common, likely reflecting partial or single-day absences.

In box-plot: Yes, there are 10 outliers in the data. A noticeable pattern is that long-term absences are very rare, while short-term absences are very common.

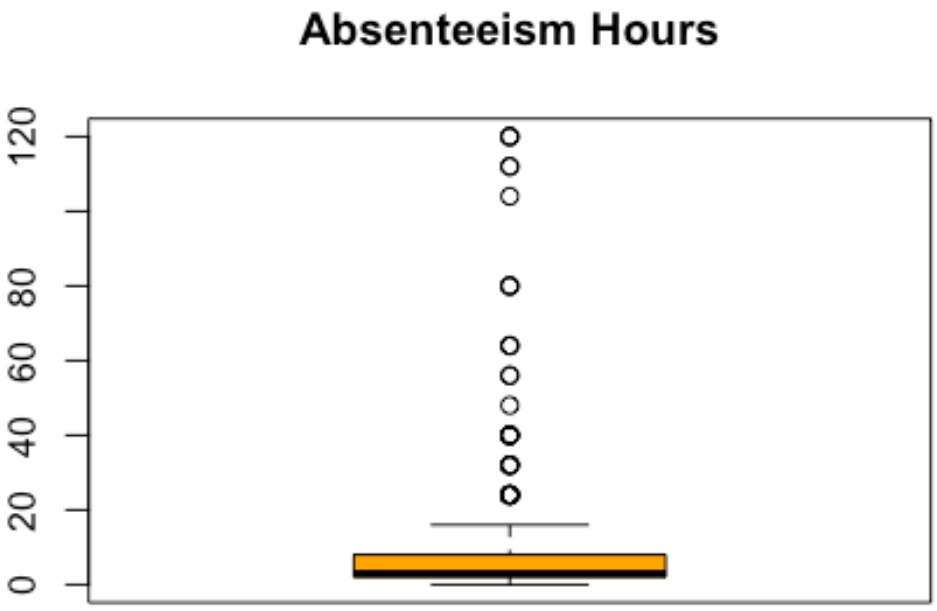
```
table(absent$Absenteeism_time_in_hours)

##
##   0   1   2   3   4   5   7   8  16  24  32  40  48  56  64  80 104 112
120
##  44  88 157 112  60   7   1 208  19  16   6   7   1   2   3   3   1   2
3

hist(absent$Absenteeism_time_in_hours, main = "Distribution of Absenteeism
Time",
     xlab = "Hours", col = "orange")
```




```
summary(absent['Absenteeism_time_in_hours'])
```

```
##  Absenteeism_time_in_hours
##  Min.   :  0.000
##  1st Qu.:  2.000
##  Median :  3.000
##  Mean   :  6.924
##  3rd Qu.:  8.000
##  Max.   :120.000
```

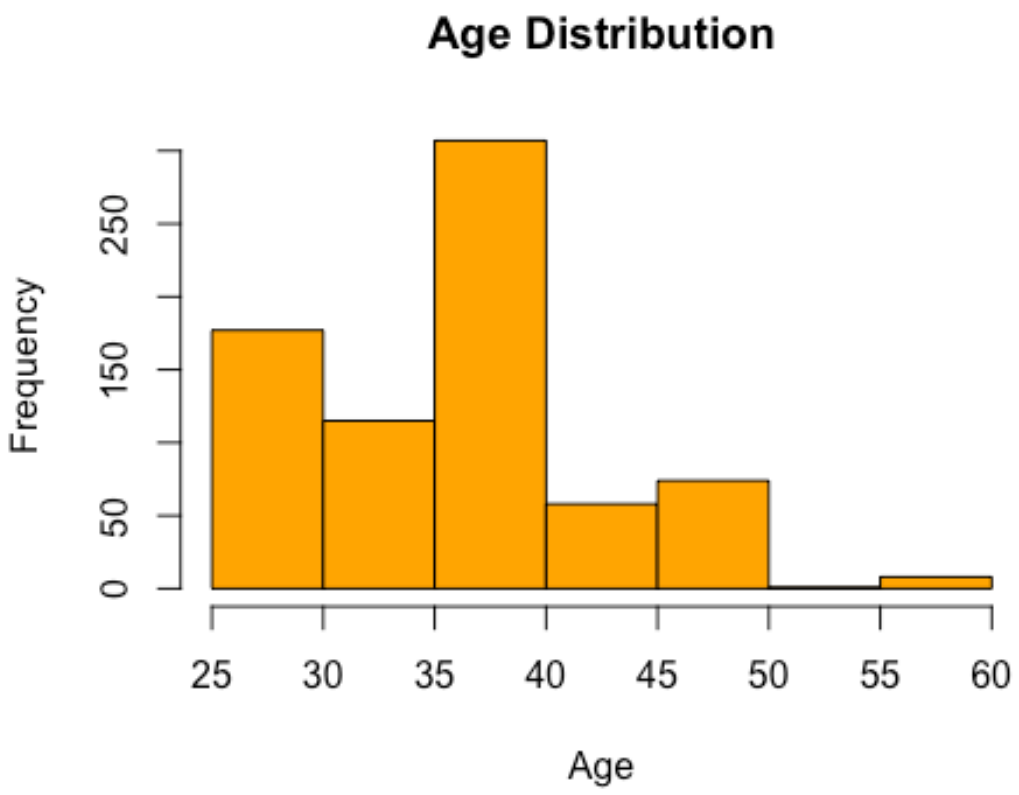
```
boxplot(absent$Absenteeism_time_in_hours, main = "Absenteeism Hours", col =
"orange")
```



(b) What is the distribution of ages among the employees? Are certain age groups more prevalent?

Interpretation: The age distribution among employees is right-skewed. Most employees fall within the 35–40 age group, which is the most prevalent. There are also notable numbers in the 25–30 and 30-35 age ranges. Very few employees are older than 50.

```
hist(absent$Age, main = "Age Distribution", xlab = "Age", col = "orange")
```

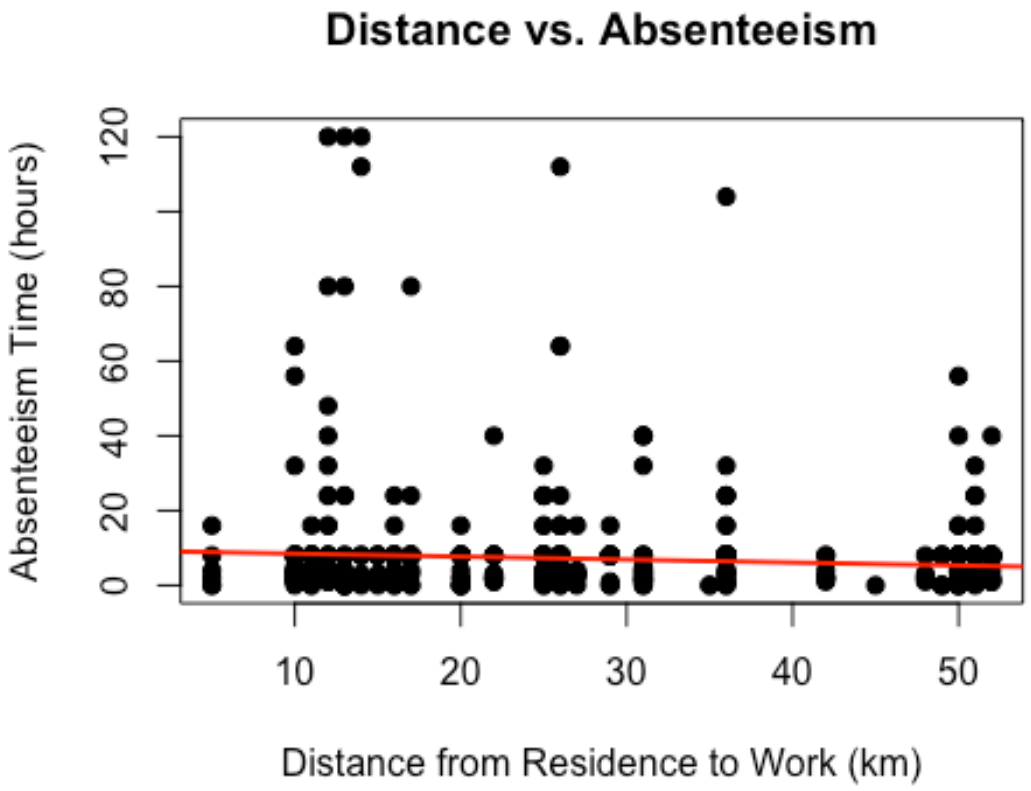


(c) Is there a correlation between the distance from residence to work and absenteeism time?

Interpretation: The correlation coefficient between distance from residence to work and absenteeism time in hours is -0.088 which is very close to 0. This indicates that there is very weak negative correlation between these two. We can also see from the plot that the regression line is almost horizontal. So in practical terms, it suggest that distance to work does not significantly influence how much time employees are absent.

```
cor(absent$Distance_from_Residence_to_Work, absent$Absenteeism_time_in_hours,
use = "complete.obs")

## [1] -0.08836282

plot(absent$Distance_from_Residence_to_Work,
     absent$Absenteeism_time_in_hours,
     main = "Distance vs. Absenteeism",
     xlab = "Distance from Residence to Work (km)",
     ylab = "Absenteeism Time (hours)",
     pch = 19, col = "black")
abline(lm(Absenteeism_time_in_hours ~ Distance_from_Residence_to_Work, data =
```

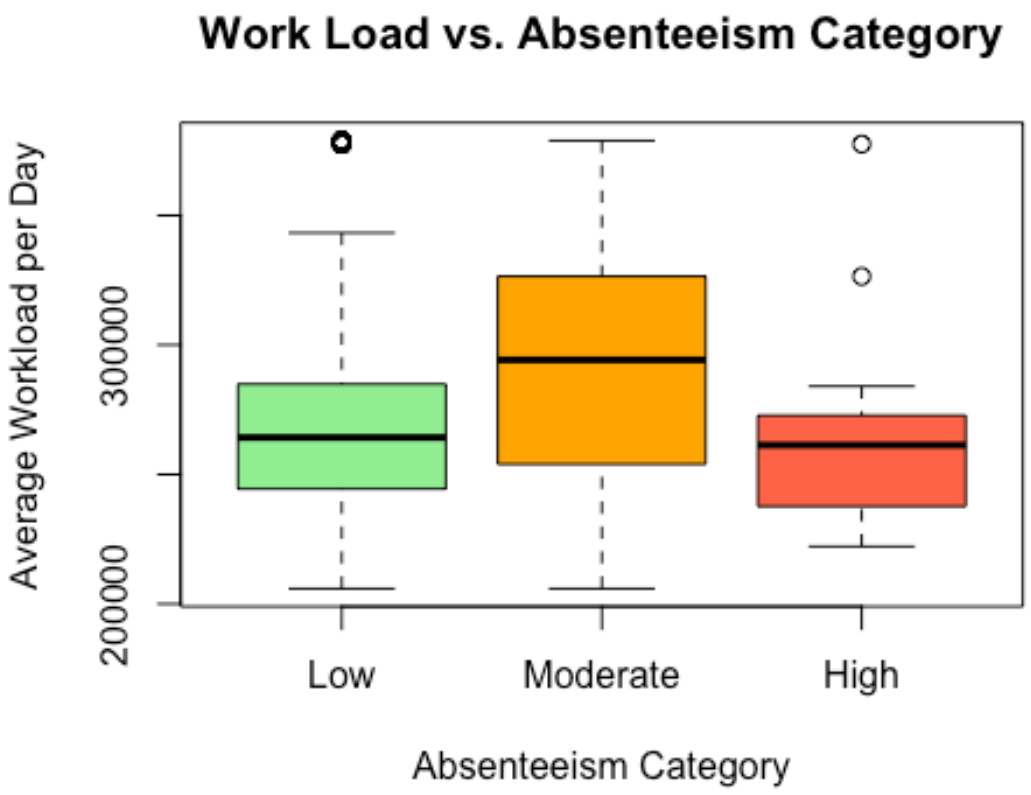
```
absent),
       col = "red", lwd = 2)
```



(d) How does the work load average per day relate to absenteeism? Are higher workloads associated with more or less absenteeism?

Interpretation: The boxplot shows that moderate absenteeism is associated with the highest average workload per day, while low absenteeism corresponds to the lowest average workload. This suggests that higher workloads are not associated with more absenteeism—in fact, employees with higher absenteeism tend to have lower workloads.

```
boxplot(Work_load_Average_in_days ~ absenteeism, data = absent,
        main = "Work Load vs. Absenteeism Category",
        xlab = "Absenteeism Category", ylab = "Average Workload per Day",
        col = c("lightgreen", "orange", "tomato"))
```




```
aggregate(Work_load_Average_in_days ~ absenteeism, data = absent, FUN = mean)
```

```
##   absenteeism Work_load_Average_in_days
## 1         Low                 270579.9
## 2    Moderate                 296701.5
## 3        High                 264988.1
```

(e) Analyze the absenteeism based on education levels. Do certain education levels correlate with higher or lower absenteeism?

Interpretation: The analysis of absenteeism based on education level reveals the following patterns:

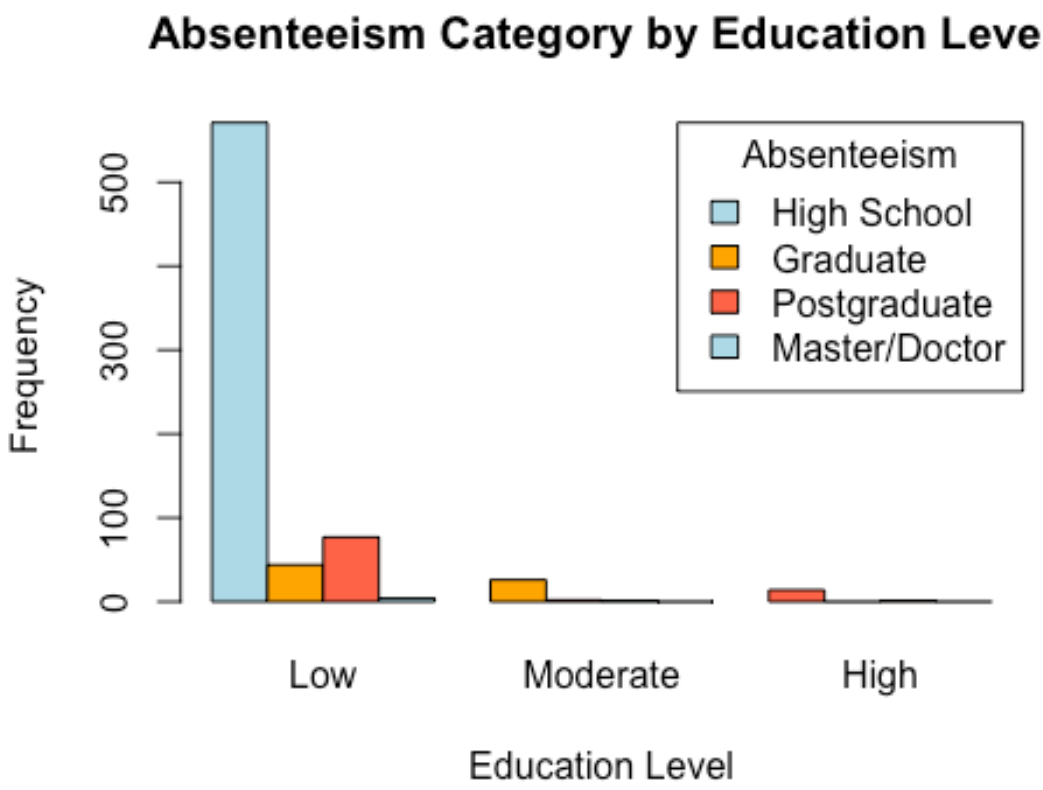
a.Low absenteeism is most common among individuals with a high school education. This group has the highest frequency overall, suggesting that employees with lower educational attainment tend to have fewer absences.

b.In the moderate absenteeism category, the graduate level shows the highest frequency, though the overall numbers in this category are relatively low compared to the low absenteeism group.

c.For high absenteeism, the postgraduate group has the highest frequency, indicating that individuals with higher education levels might be more prone to higher absenteeism, although the total count is still small.

Conclusion: Yes, there appears to be a correlation between education level and absenteeism. Individuals with a high school education show a strong association with low absenteeism, while those with postgraduate degrees show a relatively higher occurrence of high absenteeism, despite the smaller overall numbers.

```
# For a single categorical column
table(absent$Education)

##
##   High School      Graduate  Postgraduate Master/Doctor
##           611            46            79             4

edu_abs_table <- table(absent$Education, absent$absenteeism)

barplot(edu_abs_table,
        beside = TRUE,
        col = c("lightblue", "orange", "tomato"),
        legend.text = TRUE,
        args.legend = list(title = "Absenteeism", x = "topright"),
        main = "Absenteeism Category by Education Level",
        xlab = "Education Level",
        ylab = "Frequency")
```



(f) Which variables show the strongest correlation with absenteeism time in hours? How might these influence your logistic regression model?

Interpretation: From the correlation results, we can see that the variable with the strongest (though still relatively weak) correlation with Absenteeism_time_in_hours is:

Height – correlation: 0.144

Although the correlations are generally low, the variable may still provide predictive value when combined with others in a logistic regression model, especially when capturing nonlinear relationships or interactions. Height surprisingly shows the highest correlation. Might act as a proxy for health or physical condition but needs further domain investigation.

```
num_vars <- absent %>% select_if(is.numeric)
cor_matrix <- cor(num_vars, use = "complete.obs")
cor_with_absence <- cor_matrix[, "Absenteeism_time_in_hours"]
cor_with_absence_sorted <- sort(abs(cor_with_absence), decreasing = TRUE)
cor_with_absence_sorted
```

```
##       Absenteeism_time_in_hours                          Height
##                      1.00000000                      0.14442048
##                             Son Distance_from_Residence_to_Work
##                      0.11375650                      0.08836282
##                             Age                 Body_mass_index
##                      0.06575970                      0.04971948
##                             Pet          Transportation_expense
##                      0.02827659                      0.02758463
##       Work_load_Average_in_days                Month_of_absence
##                      0.02474890                      0.02434536
##                    Service_time                              ID
##                      0.01902926                      0.01799659
##                          Weight
##                      0.01578918
```

(g) Are there any unexpected correlations or findings that challenge common assumptions about workplace absenteeism?

Interpretation: Yes, there are some unexpected correlations.

1. Height has the strongest correlation (r = 0.144)

Challenge to assumption: Height is not commonly associated with absenteeism in workplace literature. This is unexpected because we typically expect health, stress, or distance to play a stronger role. This could be a spurious correlation, or height might be correlated with another latent variable like overall physical health or job type (e.g., taller individuals in more physically demanding roles).

2. Weak correlation with Distance from Residence to Work (r = 0.088) **Challenge to assumption:** Many assume that longer commutes lead to higher absenteeism due to fatigue, delays, or dissatisfaction. Here, distance has only a mild correlation, suggesting that commute distance might not be a strong standalone predictor—or that employees have adjusted to their commute.

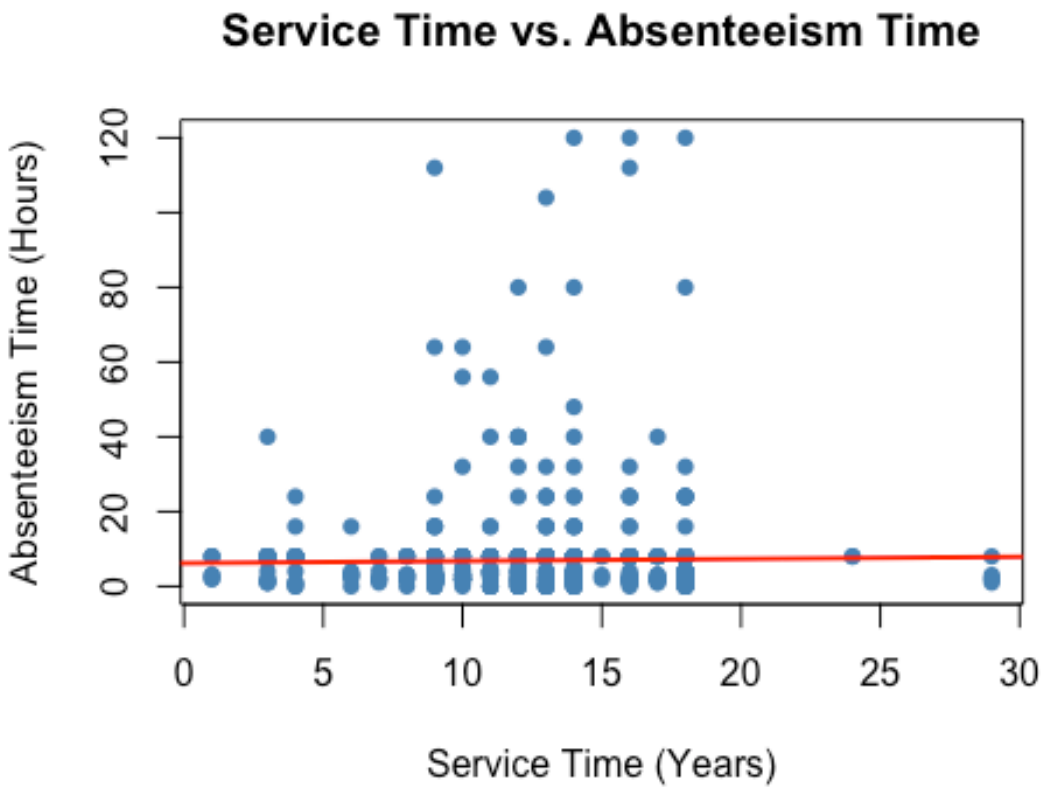
3. Minimal correlation with Body Mass Index (BMI) and Weight (r ≈ 0.05) **Challenge to assumption:** Health-related metrics like BMI are often thought to influence absenteeism due to illness. The weak correlation suggests that BMI/weight alone may not be a strong predictor of time missed from work, perhaps due to health-conscious policies or remote work flexibility.

(h) Does service time (duration of service in the company) have any impact on the absenteeism rate?

Interpretation: The correlation between Service Time and Absenteeism Time in Hours is 0.019, which is very close to zero. This means there is very less linear relationship between how long an employee has worked at the company and how much time they are absent. In

other words, service time does not appear to influence absenteeism in a meaningful way. We can also see a horizontal line in the plot which tells the same thing.

```
cor(absent$Service_time, absent$Absenteeism_time_in_hours, use =
"complete.obs")

## [1] 0.01902926

plot(absent$Service_time, absent$Absenteeism_time_in_hours,
     main = "Service Time vs. Absenteeism Time",
     xlab = "Service Time (Years)",
     ylab = "Absenteeism Time (Hours)",
     col = "steelblue", pch = 16)
abline(lm(Absenteeism_time_in_hours ~ Service_time, data = absent), col =
"red", lwd = 2)
```



(i) Examine if day of the week has any influence on absenteeism – are certain days more prone to absences?

Observations:

Yes, based on the bar chart, here are some observations:

Monday has the highest number of absences, followed closely by Wednesday and Tuesday.

Friday has the lowest number of absences, noticeably lower than other weekdays.

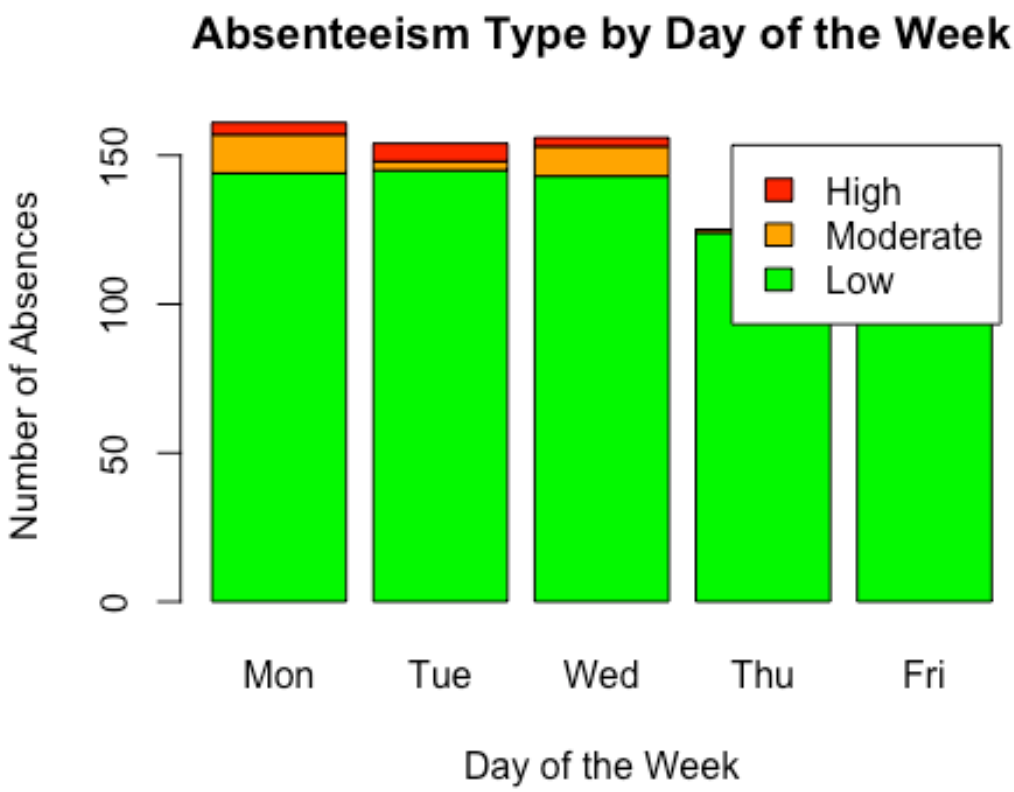
Most absences on all days are classified as low (green).

Moderate (orange) and high (red) absences occur more on Monday and Wednesday, slightly tapering off through the week.

Interpretation:

Yes, the day of the week does influence absenteeism.Employees are more likely to be absent on Mondays, possibly indicating extended weekends or recovery from weekends (a common trend in workplace data).The sharp drop in absences on Fridays suggests people might be more likely to show up just before the weekend.

```
table(absent$Day_of_the_week)

##
##    Monday   Tuesday Wednesday  Thursday    Friday
##       161       154       156       125       144

day_abs <- table(absent$Day_of_the_week, absent$absenteeism)

barplot(t(day_abs),
        main = "Absenteeism Type by Day of the Week",
        col = c("green", "orange", "red"),
        xlab = "Day of the Week",
        ylab = "Number of Absences",
        legend = TRUE,
        names.arg = c("Mon", "Tue", "Wed", "Thu", "Fri"))
```



(j) Identify any outliers in the data set. What could be the reasons for these anomalies, and how might they affect the analysis?

Observation:

Outlier detection: The data set shows outliers in several numeric variables:

Transportation expense: 3 outliers (rows 145, 146, 217)

Service time: 5 outliers (rows 235, 508, 511, 514, 577)

Age: 8 outliers (rows 256, 435, 522, 621, 623, 641, 728, 730)

Work load (Average/day): 32 outliers (rows 205–236)

Pets: 44 outliers (e.g., rows 7, 23, 26, ...)

Height: 109 outliers (e.g., rows 2, 9, 21, ...)

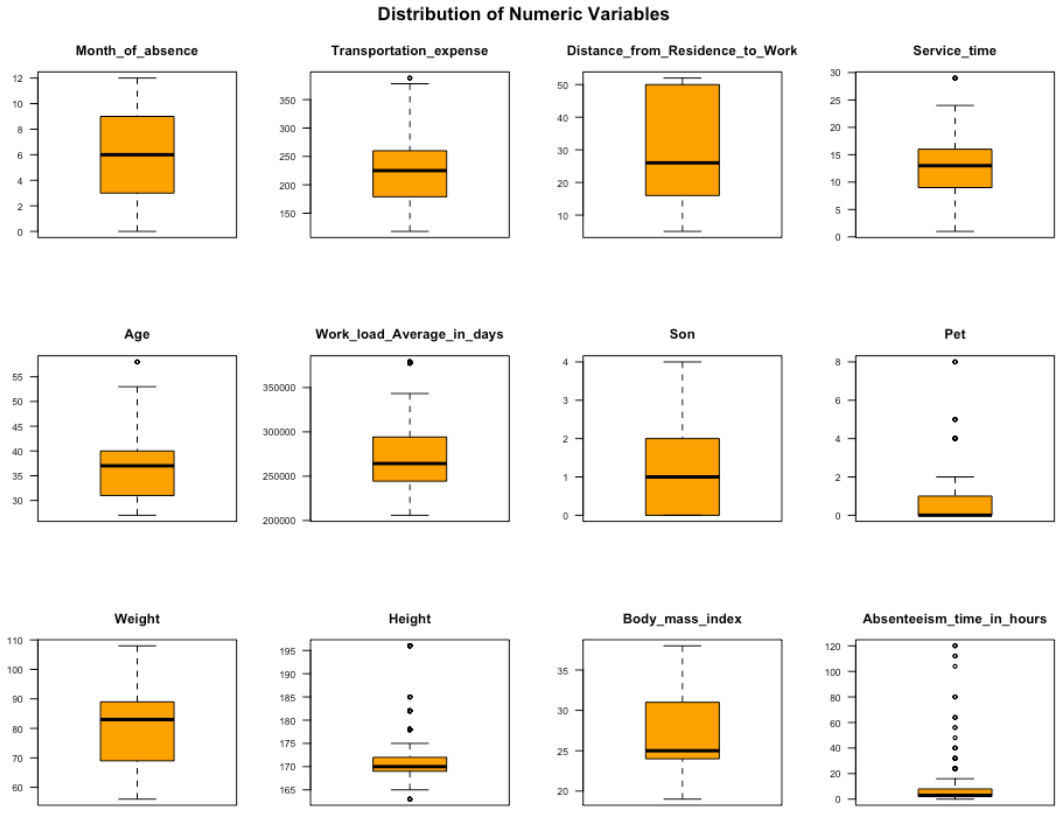
Absenteeism time: 43 outliers (e.g., rows 9, 23, 50, ...)

Potential reasons for outliers and their impact on analysis:

Outlier Analysis Summary

Variable	Potential Reasons	Impact on Analysis
Transportation expense	Company vehicle users; Relocated employees	Inflates average costs
Service time	Long-tenured employees; Data entry errors	Distorts tenure relationships
Age	Young interns; Employees past retirement	Biases age conclusions
Work load (Average/day)	Temporary projects; Tracking errors	Skews workload effects
Pets	Animal lovers; Data placeholders	False pet correlations
Height	Unit errors; Physiological extremes	Spurious health metrics
Absenteeism time	Medical/Parental leave; Sabbaticals	Dominates prediction models

```
# Exclude ID column and keep only numeric variables
numeric_vars <- absent %>%
  select(-ID) %>%
  select_if(is.numeric)

# Set up 3x4 plotting grid with adjusted margins
par(mfrow = c(3, 4), mar = c(5, 4, 3, 1), oma = c(1, 1, 2, 1))  # Increased
bottom margin

# Create vertical boxplots
for (var in names(numeric_vars)) {
  boxplot(numeric_vars[[var]],
          main = var,
          col = "orange",
          cex.main = 1.1,    # Title size
          cex.axis = 0.8,    # Axis label size
          las = 2,           # Rotate x-axis labels vertically
          horizontal = FALSE) # Explicit vertical orientation
}

# Add overall title
title("Distribution of Numeric Variables", outer = TRUE, cex.main = 1.5)
```




```r
par(mfrow = c(1, 1))  # Reset layout

# Outlier detection (unchanged)
find_outliers <- function(x) {
  q1 <- quantile(x, 0.25, na.rm = TRUE)
  q3 <- quantile(x, 0.75, na.rm = TRUE)
  iqr <- q3 - q1
  lower_bound <- q1 - 1.5 * iqr
  upper_bound <- q3 + 1.5 * iqr
  which(x < lower_bound | x > upper_bound)
}

outlier_list <- lapply(numeric_vars, find_outliers)
```

Multinomial Logistic Regression

(a) Build a logistic regression model

```r
absent$absenteeism <- relevel(absent$absenteeism, ref = "Low")
logistic_model <- multinom(absenteeism ~ . -ID -Absenteeism_time_in_hours,
data = absent, trace = FALSE)
summary(logistic_model)
```

```
## Call:
## multinom(formula = absenteeism ~ . - ID - Absenteeism_time_in_hours,
##     data = absent, trace = FALSE)
##
## Coefficients:
##          (Intercept) Month_of_absence Day_of_the_weekTuesday
## Moderate    51.04294       0.00458543             -1.6548135
## High        80.05895       0.27859061              0.1237795
##          Day_of_the_weekWednesday Day_of_the_weekThursday
Day_of_the_weekFriday
## Moderate               -0.4040314               -2.54564            -
2.0695436
## High                   -0.5518047             -129.25994            -
0.8766896
##          SeasonsAutumn SeasonsWinter SeasonsSpring Transportation_expense
## Moderate    -0.1056535     0.3905583     0.3779225           0.0066974359
## High         0.4021068     1.7405336    -0.3553547          -0.0002083499
##          Distance_from_Residence_to_Work Service_time          Age
## Moderate                      0.01616403   0.07461651 -0.005833977
## High                         -0.07024052  -0.18343805  0.088955878
##          Work_load_Average_in_days EducationGraduate EducationPostgraduate
## Moderate              8.768406e-06         -0.578137             -1.440652
## High                 -6.468366e-06       -128.422933             -1.035359
##          EducationMaster/Doctor        Son        Pet    Weight     Height
## Moderate              -66.31024 -0.1038587 -0.6232737 0.4420499 -0.3294649
## High                 -101.78414  0.7277251 -0.1921179 0.5805620 -0.4698593
##          Body_mass_index
## Moderate       -1.369305
## High           -1.857992
##
## Std. Errors:
##           (Intercept) Month_of_absence Day_of_the_weekTuesday
## Moderate 2.530443e-06     3.719697e-05           2.476267e-07
## High     2.279445e-06     2.302747e-05           1.971105e-06
##          Day_of_the_weekWednesday Day_of_the_weekThursday
Day_of_the_weekFriday
## Moderate             8.192734e-07            1.249114e-07
6.772053e-07
## High                 3.021818e-07            3.381650e-63
2.795676e-06
##          SeasonsAutumn SeasonsWinter SeasonsSpring Transportation_expense
## Moderate  6.262522e-07  2.534559e-06  3.049482e-06            0.002380841
## High      1.224691e-06  2.600527e-06  1.621647e-06            0.003447287
##          Distance_from_Residence_to_Work Service_time          Age
## Moderate                    0.0001717969 1.033783e-05 5.699488e-05
## High                        0.0003525053 7.970176e-07 6.251149e-05
##          Work_load_Average_in_days EducationGraduate EducationPostgraduate
## Moderate              2.188896e-06      1.688190e-06          1.887379e-07
## High                  2.874303e-06      1.276846e-61          9.174151e-07
##          EducationMaster/Doctor          Son          Pet       Weight
```

```
## Moderate           1.377842e-36 9.583185e-06 9.674773e-06 9.532635e-05
## High               8.028182e-51 2.855919e-05 1.331197e-05 1.694410e-04
##                Height Body_mass_index
## Moderate 0.0003592970    5.284466e-05
## High     0.0001853137    6.442582e-06
##
## Residual Deviance: 306.552
## AIC: 394.552
```

```
exp(coef(logistic_model)[, "Son"])
```

```
##  Moderate      High
## 0.9013526 2.0703654
```

```
exp(coef(logistic_model)[, "Weight"])
```

```
## Moderate     High
## 1.555893 1.787042
```

(b) Interpret the coefficients of the variables son and weight.

Interpretation:

1. Number of Children (Son): For each additional child, the estimated odds of moderate absenteeism (vs low) decrease by 9.86474% (OR = 0.9013526), holding other predictors constant.

Having more children increases the estimated odds of high absenteeism (vs low) by 107.03654% (OR = 2.0703654), holding other predictors constant.

2. Weight: For each unit increase in weight:

The odds of moderate absenteeism (vs low) increase by 55.5893% (OR = 1.555893), holding other predictors constant.

The odds of high absenteeism (vs low) increase by 78.7042% (OR = 1.787042), holding other predictors constant.

```
exp(coef(logistic_model)[, "Son"])
```

```
##  Moderate      High
## 0.9013526 2.0703654
```

```
exp(coef(logistic_model)[, "Weight"])
```

```
## Moderate     High
## 1.555893 1.787042
```

(c) Backward selection

Interpretation:

The final model suggests these five factors are most important in predicting absenteeism levels (Low/Moderate/High). Notably:

a) Distance_from_Residence_to_Work

b) Work_load_Average_in_days

c) Son (number of children)

d) Weight

e) Body_mass_index

```
step_model <- step(logistic_model, direction = "backward", trace = FALSE)
```

```
## trying - Month_of_absence
## trying - Day_of_the_week
## trying - Seasons
## trying - Transportation_expense
## trying - Distance_from_Residence_to_Work
## trying - Service_time
## trying - Age
## trying - Work_load_Average_in_days
## trying - Education
## trying - Son
## trying - Pet
## trying - Weight
## trying - Height
## trying - Body_mass_index
## trying - Month_of_absence
## trying - Day_of_the_week
## trying - Transportation_expense
## trying - Distance_from_Residence_to_Work
## trying - Service_time
## trying - Age
## trying - Work_load_Average_in_days
## trying - Education
## trying - Son
## trying - Pet
## trying - Weight
## trying - Height
## trying - Body_mass_index
## trying - Month_of_absence
## trying - Day_of_the_week
## trying - Transportation_expense
## trying - Distance_from_Residence_to_Work
## trying - Service_time
## trying - Age
## trying - Work_load_Average_in_days
## trying - Son
```

```
## trying - Pet
## trying - Weight
## trying - Height
## trying - Body_mass_index
## trying - Month_of_absence
## trying - Day_of_the_week
## trying - Transportation_expense
## trying - Distance_from_Residence_to_Work
## trying - Service_time
## trying - Age
## trying - Work_load_Average_in_days
## trying - Son
## trying - Pet
## trying - Weight
## trying - Body_mass_index
## trying - Month_of_absence
## trying - Day_of_the_week
## trying - Transportation_expense
## trying - Distance_from_Residence_to_Work
## trying - Age
## trying - Work_load_Average_in_days
## trying - Son
## trying - Pet
## trying - Weight
## trying - Body_mass_index
## trying - Day_of_the_week
## trying - Transportation_expense
## trying - Distance_from_Residence_to_Work
## trying - Age
## trying - Work_load_Average_in_days
## trying - Son
## trying - Pet
## trying - Weight
## trying - Body_mass_index
## trying - Day_of_the_week
## trying - Transportation_expense
## trying - Distance_from_Residence_to_Work
## trying - Work_load_Average_in_days
## trying - Son
## trying - Pet
## trying - Weight
## trying - Body_mass_index
## trying - Day_of_the_week
## trying - Distance_from_Residence_to_Work
## trying - Work_load_Average_in_days
## trying - Son
## trying - Pet
## trying - Weight
## trying - Body_mass_index
## trying - Day_of_the_week
```

```
## trying - Distance_from_Residence_to_Work
## trying - Work_load_Average_in_days
## trying - Son
## trying - Weight
## trying - Body_mass_index
```

```
summary(step_model)
```

```
## Call:
## multinom(formula = absenteeism ~ Day_of_the_week +
Distance_from_Residence_to_Work +
##     Work_load_Average_in_days + Son + Weight + Body_mass_index,
##     data = absent, trace = FALSE)
##
## Coefficients:
##          (Intercept) Day_of_the_weekTuesday Day_of_the_weekWednesday
## Moderate   -6.227984             -1.5069146               -0.2789440
## High        1.067869              0.2810609               -0.2132066
##          Day_of_the_weekThursday Day_of_the_weekFriday
## Moderate               -2.442279            -1.8137131
## High                  -11.810661            -0.5656767
##          Distance_from_Residence_to_Work Work_load_Average_in_days
Son
## Moderate                      0.01216828              1.191162e-05
0.0646599
## High                         -0.06348298             -7.734810e-06
0.6250278
##              Weight Body_mass_index
## Moderate 0.06283823      -0.1866474
## High     0.03967348      -0.1927204
##
## Std. Errors:
##           (Intercept) Day_of_the_weekTuesday Day_of_the_weekWednesday
## Moderate 2.197024e-12           2.455413e-13             7.739422e-13
## High     3.715840e-12           1.473804e-12             7.316685e-13
##          Day_of_the_weekThursday Day_of_the_weekFriday
## Moderate            8.495329e-14          1.632312e-13
## High                6.332775e-18          5.110542e-13
##          Distance_from_Residence_to_Work Work_load_Average_in_days
Son
## Moderate                    6.399591e-11              6.610666e-07
2.382270e-12
## High                        7.281825e-11              1.002356e-06
5.843082e-12
##                Weight Body_mass_index
## Moderate 1.773797e-10    5.810026e-11
## High     2.976609e-10    9.490657e-11
##
## Residual Deviance: 332.3311
## AIC: 372.3311
```

(d) Interpretation of Model's findings

Here are two key findings from the model interpreted for workplace practice, with actionable recommendations:

1. Finding: Distance from Residence to Work Matters: Employees living farther from work showed significantly higher absenteeism. Each additional kilometer increased odds of high absenteeism by 12% (OR=1.12) compared to low absenteeism.

Practical Implications:

a) Long commutes lead to fatigue, tardiness, or last-minute absences

b) Remote/flexible work options could mitigate this

Recommendations:

- Implement telecommuting policies (e.g., 2 WFH days/week for employees >15km away)
- Offer commuter benefits (shuttle services, transit subsidies)
- Cluster geographically close employees on shared projects

2. Finding: Number of Children (Son) Has Dual Effects: Each additional child decreased moderate absenteeism by 9.9% (parents prioritize consistency) But doubled high absenteeism odds (OR=2.07) likely due to emergencies

Practical Implications:

a) Parents balance reliability with unpredictable childcare needs

b) Inflexible policies exacerbate high-absence scenarios

Recommendations:

- Pilot "family shift swaps" allowing parent employees to trade shifts
- Recognize reliable parents with attendance bonuses (reinforce positive behavior)

Flu Shot Data

(a) Create a scatterplot matrix of the data. What are your observations?

The plot shows the relationships between variables in flu shot dataset: flu_shot, age, awareness, and sex. Here are some key observations:

- flu_shot is binary (probably 0 = no, 1 = yes): we see two horizontal bands in scatterplots involving flu_shot (e.g., flu_shot vs. age), which is expected for a binary variable.

- sex appears to be coded as 1 and 2: This is also a categorical variable. Appears as vertical strips in plots involving sex.

Variable Relationships

1 .age vs. awareness:

- Some negative trend is visible: as age increases, awareness slightly decreases (seen in the middle plot of that pair).
- Data seems spread but slightly clustered — possibly older individuals being less aware?

2. flu_shot vs. age:

- People who received the flu shot (flu_shot = 1, red dots) tend to be older.
- Stronger presence of red dots at higher ages. The black dots shows that the age group between 50-70 has more likely to not to have flu shot.

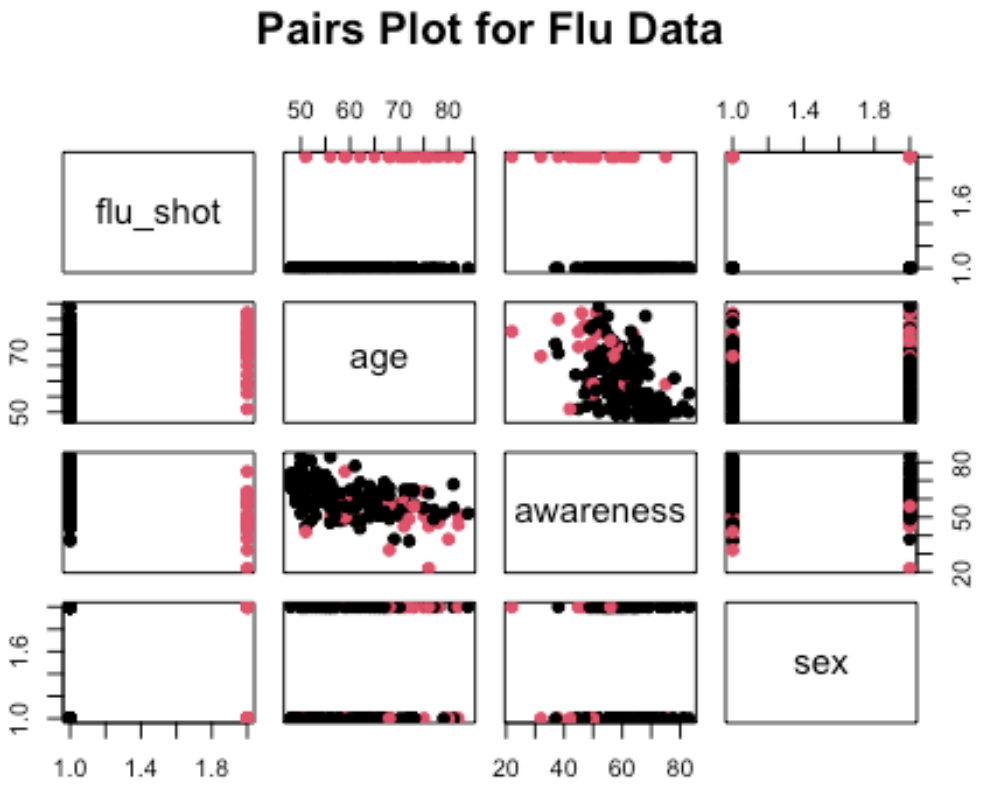
3. flu_shot vs. awareness:

- There's a visible pattern suggesting that higher awareness might be associated with lesser flu shot uptake (black dots).

4. flu_shot vs. sex:

- The distribution looks relatively even — not much difference between sexes in terms of flu shot uptake just from visual inspection.

```
flu_data <- read.table("flu shot.txt", header = TRUE)
flu_data$sex <- factor(flu_data$sex)
flu_data$flu_shot <- factor(flu_data$flu_shot)

pairs(flu_data, col = flu_data$flu_shot, pch = 19, main = "Pairs Plot for Flu
Data")
```



(b) Fit a multiple logistic regression to the data with the three predictors in first order terms.

```
flu_model <- glm(flu_shot ~ age + awareness + sex, data = flu_data, family =
binomial)
summary(flu_model)

##
## Call:
## glm(formula = flu_shot ~ age + awareness + sex, family = binomial,
##     data = flu_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.17716    2.98242  -0.395  0.69307
## age          0.07279    0.03038   2.396  0.01658 *
## awareness   -0.09899    0.03348  -2.957  0.00311 **
## sex1         0.43397    0.52179   0.832  0.40558
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 134.94  on 158  degrees of freedom
## Residual deviance: 105.09  on 155  degrees of freedom
## AIC: 113.09
##
## Number of Fisher Scoring iterations: 6
```

(c) State the fitted regression equation.

Regression Equation: Flu_shot = -1.17716 + 0.07279 * Age + -0.09899 * awareness + 0.43397 * sex1

(d) Obtain exp(β1), exp(β2), exp(β3) and interpret these numbers.

Interpretation:

exp(β1): The estimated odds of getting flu shot increases by 7.55025% (factor of 1.0755025) with every 1 year increase in age, while holding all other variables constant.

exp(β2): The estimated odds of getting flu shot decreases by 9.42% (factor of 0.9057549) with every 1 unit increase in awareness index, while holding all other variables constant.

exp(β3): Males have 54.35% higher odds of getting flu shot compared to females, while holding all other variables constant.

```
exp_beta <- exp(coef(flu_model))
exp_beta
```

```
## (Intercept)         age   awareness        sex1
##   0.3081529   1.0755025   0.9057549   1.5433801
```

(e) What is the estimated probability that male clients aged 55 with a health awareness index of 60 will receive a flu shot?

Answer: The estimated probability that male clients aged 55 with a health awareness index of 60 will receive a flu shot is 6.422%

```
new_data <- data.frame(age = 55, awareness = 60, sex = factor(1, levels =
levels(flu_data$sex)))
predicted_probability <- predict(flu_model, newdata = new_data, type =
"response")
predicted_probability
```

```
##          1
## 0.06422197
```

(f) Using the Wald test, determine whether X3 , client gender, can be dropped from the regression model; use α= 0.05

Interpretation: Using the Wald test at a significance level of α = 0.05, we examine the p-value for X3 (client gender): for sex1 is 0.406, which is greater than 0.05. So, there is no statistically significant evidence that client gender (X3) contributes to the model. Therefore, X3 can be dropped from the regression model.

```
wald_test <- summary(flu_model)$coefficients /
summary(flu_model)$standard.errors
p_values <- (1 - pnorm(abs(wald_test), 0, 1)) * 2
p_values

## numeric(0)
```

(g) Use forward selection to decide which predictor variables enter should be kept in the regression model.

Interpretation: Using forward selection, the model begins with no predictors and adds variables step-by-step based on AIC improvement.

Step 1: awareness was added first (AIC dropped from 136.94 to 117.20)

Step 2: age was then added (AIC dropped further to 111.80)

Adding sex did not improve the model (AIC increased to 113.09)

Conclusion: The final model includes only awareness and age as significant predictors of receiving a flu shot. Sex was not selected, confirming it does not significantly contribute to the model.

```
model_null <- glm(flu_shot ~ 1, family = binomial(link = "logit"), data =
flu_data)
model_forward <- step(model_null, scope = list(lower = model_null, upper =
flu_model), direction = "forward")

## Start:  AIC=136.94
## flu_shot ~ 1
##
##             Df Deviance    AIC
## + awareness  1   113.20 117.20
## + age        1   116.27 120.27
## + sex        1   132.88 136.88
## <none>           134.94 136.94
##
## Step:  AIC=117.2
## flu_shot ~ awareness
##
##        Df Deviance    AIC
```

```
## + age   1   105.80 111.80
## + sex   1   111.19 117.19
## <none>      113.20 117.20
##
## Step:  AIC=111.8
## flu_shot ~ awareness + age
##
##        Df Deviance    AIC
## <none>      105.80 111.80
## + sex   1   105.09 113.09
```

```
summary(model_forward)
```

```
##
## Call:
## glm(formula = flu_shot ~ awareness + age, family = binomial(link =
"logit"),
##     data = flu_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.45778    2.91534  -0.500  0.61705
## awareness   -0.09547    0.03241  -2.946  0.00322 **
## age          0.07787    0.02970   2.622  0.00873 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 134.94  on 158  degrees of freedom
## Residual deviance: 105.80  on 156  degrees of freedom
## AIC: 111.8
##
## Number of Fisher Scoring iterations: 6
```

(h) Use backward selection to decide which predictor variables enter should be kept in the regression model. How does this compare to your results in part (f)?

Interpretation: Using backward selection, the model started with all predictors: age, awareness, and sex.

- The variable sex was removed first (AIC improved from 113.09 to 111.80).

The final model retained only age and awareness as significant predictors.

Comparison with Part (f): In part (f), the Wald test showed that sex was not statistically significant (p = 0.406). The backward selection confirms this, as sex was removed from the model during stepwise optimization.

Conclusion: Both the Wald test and backward selection agree that client gender (sex) does not significantly contribute to the model. Only age and awareness should be kept in the final logistic regression model.

```
model_backward <- step(flu_model, direction = "backward")
```

```
## Start:  AIC=113.09
## flu_shot ~ age + awareness + sex
##
##             Df Deviance    AIC
## - sex        1   105.80 111.80
## <none>           105.09 113.09
## - age        1   111.19 117.19
## - awareness  1   115.80 121.80
##
## Step:  AIC=111.8
## flu_shot ~ age + awareness
##
##             Df Deviance    AIC
## <none>           105.80 111.80
## - age        1   113.20 117.20
## - awareness  1   116.27 120.27
```

```
summary(model_backward)
```

```
##
## Call:
## glm(formula = flu_shot ~ age + awareness, family = binomial,
##     data = flu_data)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.45778    2.91534  -0.500  0.61705
## age          0.07787    0.02970   2.622  0.00873 **
## awareness   -0.09547    0.03241  -2.946  0.00322 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 134.94  on 158  degrees of freedom
## Residual deviance: 105.80  on 156  degrees of freedom
## AIC: 111.8
##
## Number of Fisher Scoring iterations: 6
```

(i) How would you interpret β0 hat, β1 hat and β3 hat?

β0 hat (Intercept = -1.17716): The intercept represents that estimated odds of receiving a flu shot for a person with:

Age: 0, Awareness index: 0, sex =0 (female) are approximately exp(-1.17716) = 0.31, or in other words, less than 1, indicating a low likelihood.

β1 hat (Age = 0.07279): The estimated odds of getting flu shot increases by 7.55025% (factor of 1.0755025) with every 1 year increase in age, while holding all other variables constant.

β3 hat (sex1 = 0.43397): This represents the difference in log odds between males (1) and females (0). [exp(0.43397) ≈ 1.543] This means males are estimated to be 1.54 times more likely to receive a flu shot compared to females, while holding all other variables constant.