Homework 2

Group-3 Members: Asha Shah | Prabhath Pasula | Rithwik Reddy Nandyala

April 14, 2025

Introduction

In this task, we discuss the problem of absenteeism in the modern dynamic workplace environment using logistic regression. We are attempting to forecast and understand the patternsof absenteeism founded on a diverse dataset of a Brazilian courier company. Also, we are studying flu shot statistics to learn the determinants of flu vaccination uptake.

Setup

```r
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
```

```r
library(readxl)
library(nnet)
library(GGally)
```

```
## Loading required package: ggplot2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```r
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.4     v tibble    3.2.1
## v purrr     1.0.4     v tidyr     1.3.1
## v readr     2.1.5

## -- 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 error
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
##
## The following object is masked from 'package:dplyr':
##
##     select
```

#______________________________ # Absenteeism Data Analysis #______________________________

Question (a): How is the ‘Absenteeism time in hours’ distributed? Are there any noticeable patterns or outliers?

Explanation Exploratory Data Analysis (EDA) involves summarizing the main characteristics of a dataset by using visual methods. This step is crucial to understand the structure of the data, identify patterns, detect outliers, and check assumptions before performing further analysis.

```
# Load and prepare the data
absent = read_excel("Absenteeism_at_work.xls")
names(absent) #column names
```

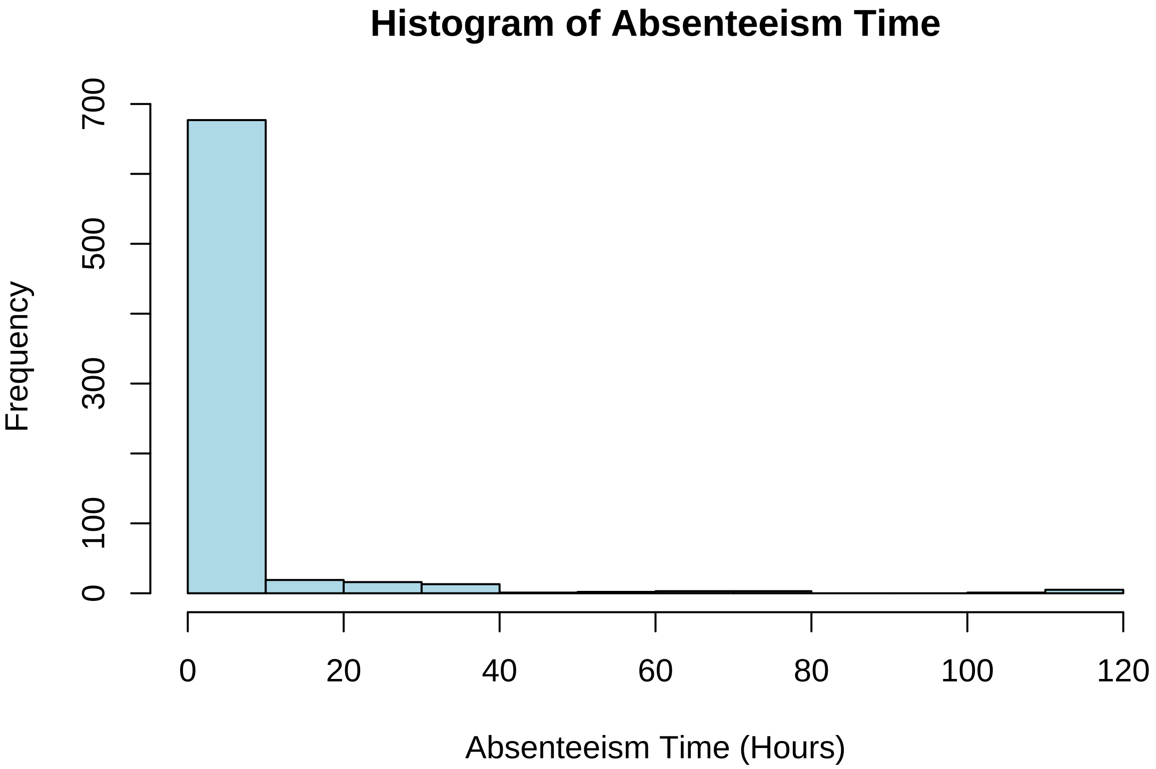
Code

```
##  [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"
```

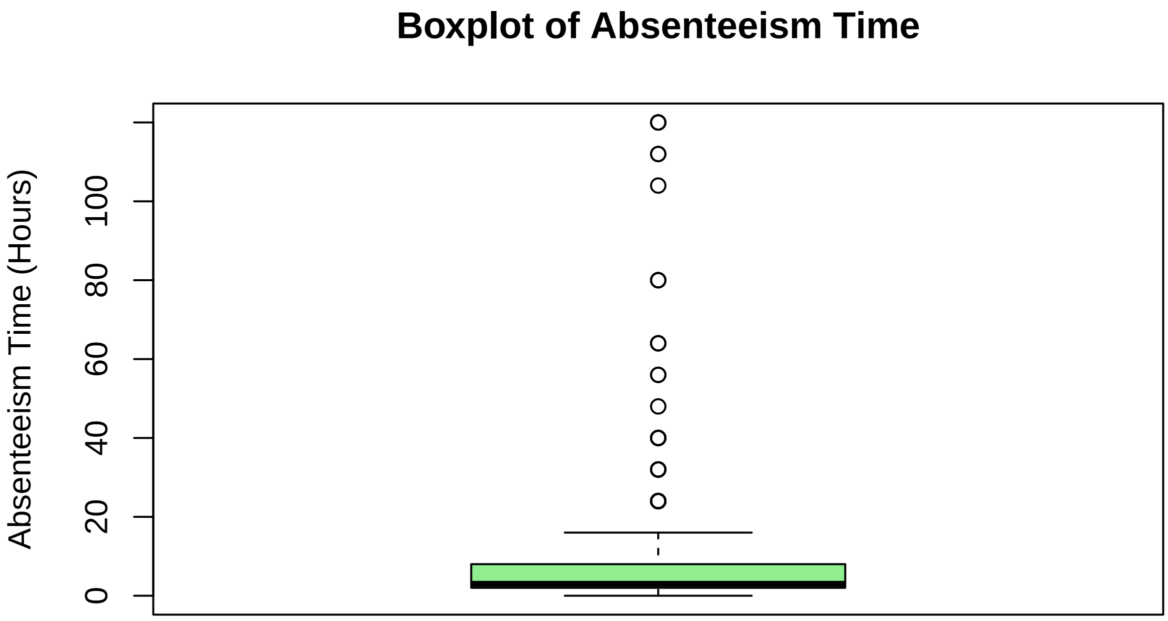
```
# Recode absenteeism into categories
absent <- absent %>%
  mutate(absenteeism = case_when(
    Absenteeism_time_in_hours >= 0 & Absenteeism_time_in_hours <= 20 ~ "Low",
    Absenteeism_time_in_hours > 20 & Absenteeism_time_in_hours <= 40 ~ "Moderate",
    Absenteeism_time_in_hours > 40 ~ "High"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


# Convert categorical variables to factors
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"))


# Summary statistics to identify patterns and outliers
summary(absent$Absenteeism_time_in_hours)
```

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

```
#Analyze the distribution of 'Absenteeism time in hours'
# Histogram
hist(absent$Absenteeism_time_in_hours,
     main = "Histogram of Absenteeism Time",
     xlab = "Absenteeism Time (Hours)",
     col = "lightblue")
```



```
# Boxplot to detect outliers
boxplot(absent$Absenteeism_time_in_hours,
        main = "Boxplot of Absenteeism Time",
        ylab = "Absenteeism Time (Hours)",
        col = "lightgreen")
```



```
The histogram of Absenteeism time in hours is highly skewed to the right, with the
bulk of workers in the category of low absenteeism (0-10 hours). The histogram shows
a strong frequency at the lower end, and the boxplot highlights a few outliers at
absenteeism higher than 10 hours, as high as 120 hours. These outliers signal
occasional cases of extended absenteeism, which deserves further investigation.
Generally, data cluster around small values, but some extreme points pull the
upper tail of the distribution upwards.
```

Observations for Question (a)

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

Explanation Understanding the distribution of a variable helps to know its spread, central tendency trend, and if there are any outliers. This is valuable information for selecting appropriate statistical methods for analysis. #### Code

```
# Summary statistics for age
summary(absent$Age)
```

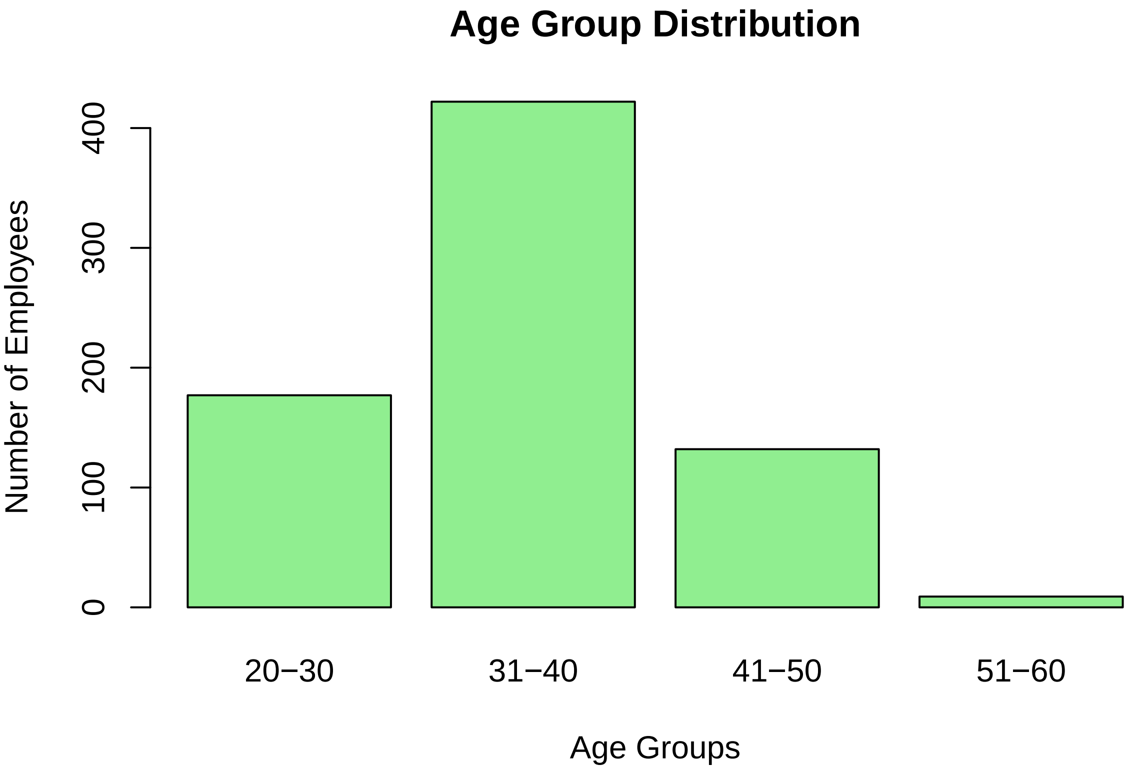
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   27.00   31.00   37.00   36.45   40.00   58.00
```

```
# Categorize ages into groups
age_groups <- absent %>%
  mutate(age_group = case_when(
    Age >= 20 & Age <= 30 ~ "20-30",
    Age >= 31 & Age <= 40 ~ "31-40",
    Age >= 41 & Age <= 50 ~ "41-50",
    Age >= 51 & Age <= 60 ~ "51-60",
    Age > 60 ~ "61+"
  )) %>%
  group_by(age_group) %>%
  summarise(count = n())
```

```
# View the frequency of age groups
print(age_groups)
```

```
## # A tibble: 4 x 2
##   age_group count
##   <chr>     <int>
## 1 20-30       177
## 2 31-40       422
## 3 41-50       132
## 4 51-60         9
```

```
# Bar chart of age groups
barplot(age_groups$count,
        names.arg = age_groups$age_group,
        main = "Age Group Distribution",
        xlab = "Age Groups",
        ylab = "Number of Employees",
        col = "lightgreen")
```



```
The distribution of employees by age shows that most of the employees belong to
the 31-40 age group, with 422 employees, the most frequent. The second is the
20-30 age group, with 177 employees, and the 41-50 age group with 132 employees.
The least frequent age group is the 51-60 age group, with a total of 9 employees.

The summary statistics reveal the minimum age of 27, the maximum of 58, and the
median of 37, which tells us that the data is centered on mid-career employees.
The age distribution is roughly symmetric with the majority of employees having
ages between 31 and 40 years, as revealed by both the bar graph and the frequency
```

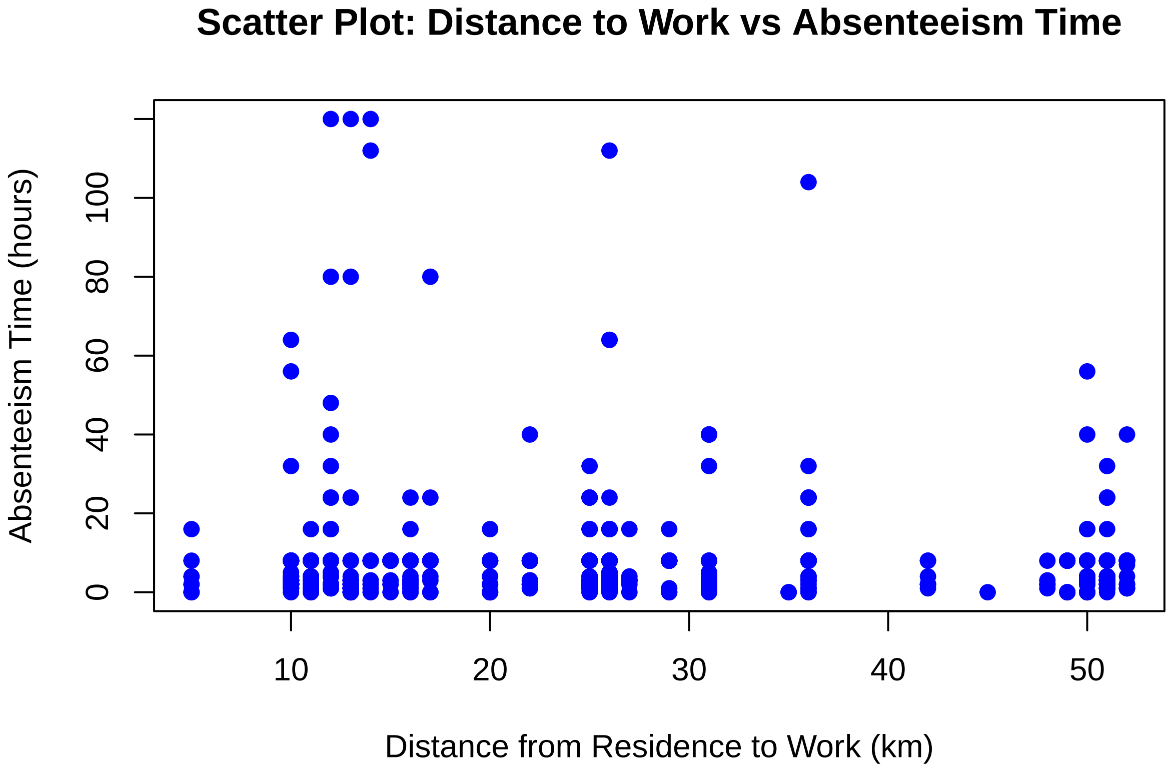
```
table. The structure of this population reveals an experienced workforce with the
majority of the employees not being too young or near retirement.
```

Observations for Question (b)

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

Explanation Understanding the correlation between home to workplace distance and absenteeism length can ascertain whether greater distances of commuting influence worker absenteeism. This can inform whether proximity to the workplace is a determining factor in workers' work attendance. 1. Use a **scatter plot** to visualize the relationship between the two variables: `Distance_from_Residence_to_Work` and `Absenteeism_time_in_hours`. 2. Calculate the **correlation coefficient** to quantify the strength and direction of the relationship. 3. Interpret the results to determine if the distance significantly impacts absenteeism.

```
# Scatter plot of distance vs absenteeism time
plot(absent$Distance_from_Residence_to_Work, absent$Absenteeism_time_in_hours,
     main = "Scatter Plot: Distance to Work vs Absenteeism Time",
     xlab = "Distance from Residence to Work (km)",
     ylab = "Absenteeism Time (hours)",
     col = "blue", pch = 19)
```



Code

```
# Calculate and display the Pearson correlation coefficient
correlation <- cor(absent$Distance_from_Residence_to_Work,
                   absent$Absenteeism_time_in_hours, method = "pearson")
correlation
```

```
## [1] -0.08836282
```

```
The scatter plot and correlation coefficient both indicate no actual correlation
between Distance from Residence to Work and Absenteeism Time. The correlation
coefficient is -0.088, and that is effectively 0, which indicates no actual linear
relationship. This would imply that commute distance has no influence on absenteeism,
and other factors may be more significant.
```

Observations for Question (c)

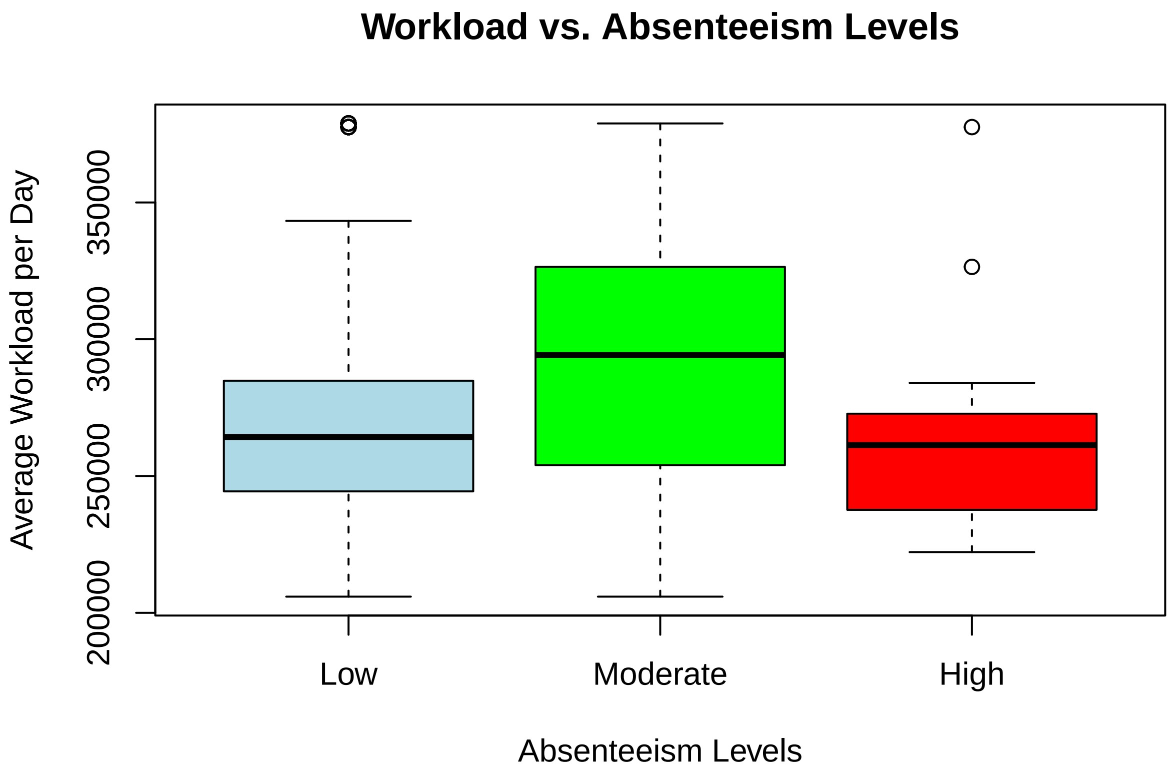
Question (d): How does the work load average per day relate to absenteeism? Are higher workloads associated with more or less absenteeism?

```r
# Summary statistics for workload average per absenteeism category
aggregate(Work_load_Average_in_days ~ absenteeism, data = absent, mean)
```

Code

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

```r
# Boxplot to visualize workload vs absenteeism levels
boxplot(Work_load_Average_in_days ~ absenteeism, data = absent,
        main = "Workload vs. Absenteeism Levels",
        xlab = "Absenteeism Levels",
        ylab = "Average Workload per Day",
        col = c("lightblue", "green", "red"))
```



```r
# ANOVA test to check if differences are statistically significant
anova_result <- aov(Work_load_Average_in_days ~ absenteeism, data = absent)
summary(anova_result)
```

```
##              Df    Sum Sq   Mean Sq F value  Pr(>F)
## absenteeism   2 1.964e+10 9.822e+09   6.535 0.00154 **
## Residuals   737 1.108e+12 1.503e+09
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Higher workloads are associated with moderate absenteeism
(highest workload:296,701.5). An ANOVA test confirms the differences in
workload across absenteeism levels are statistically significant
((p = 0.00154)). However, high absenteeism corresponds to the lowest workload,
suggesting that excessive absenteeism may not directly result from higher
workloads.
```

Observations for Question (d)

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

Explanation To analyze absenteeism by education levels, we would like to know how absenteeism varies by different education levels. This involves:

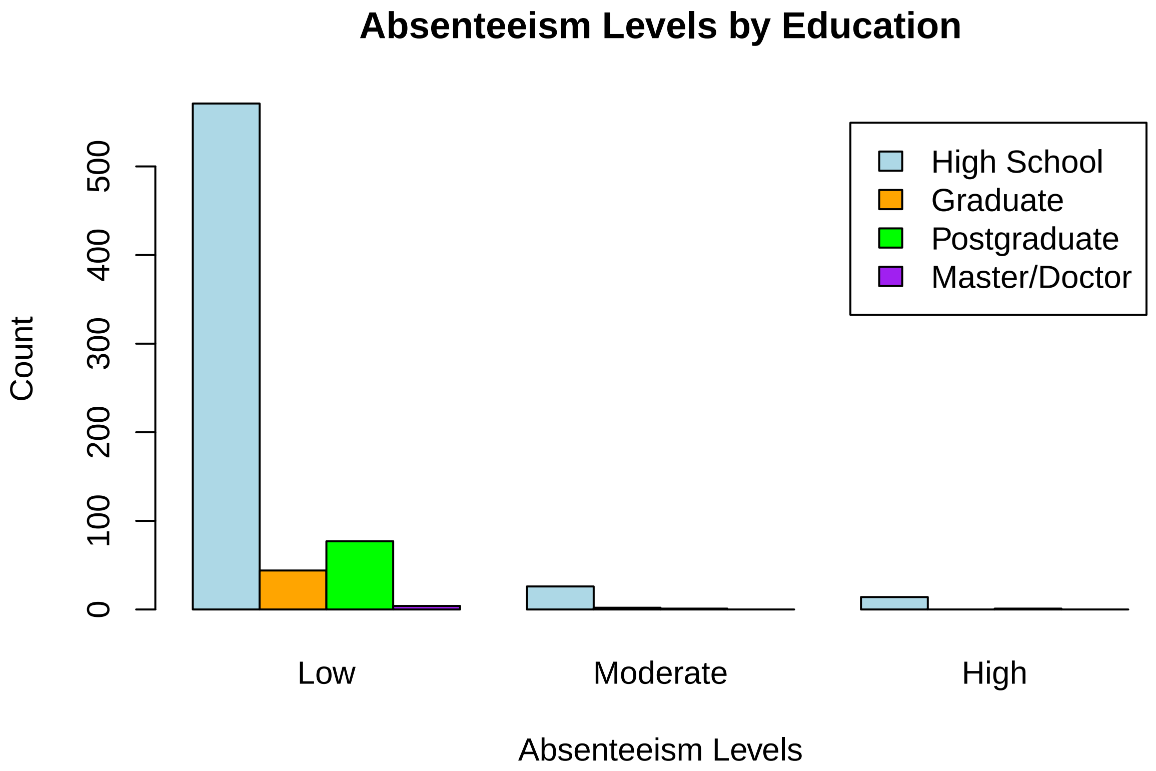
Categorizing Data: Splitting the data into education levels (e.g., High School, Graduate, Postgraduate, Doctorate). Summarizing Absenteeism: Finding the frequency or proportion of the levels of absenteeism (Low, Moderate, High) for each level of education. Visual Interpretation: Use visual tools like bar plots to visualize trends across education levels. Statistical Testing: Fisher's Exact Test is a statistical test used to determine whether there is a significant association (or relationship) between two categorical variables.

```r
# Summary statistics: Count of absenteeism levels by education
education_absenteeism <- table(absent$Education, absent$absenteeism)
print(education_absenteeism)
```

Code

```
##
##                  Low Moderate High
##   High School    571       26   14
##   Graduate        44        2    0
##   Postgraduate    77        1    1
##   Master/Doctor    4        0    0
```

```r
# Barplot to visualize absenteeism by education level
barplot(education_absenteeism, beside = TRUE,
        col = c("lightblue", "orange", "green", "purple"),
        legend = rownames(education_absenteeism),
        main = "Absenteeism Levels by Education",
        xlab = "Absenteeism Levels",
        ylab = "Count")
```



```
fisher_result <- fisher.test(education_absenteeism)
print(fisher_result)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  education_absenteeism
## p-value = 0.8036
## alternative hypothesis: two.sided
```

```

High school education dominates absenteeism, especially in the Low category,
while higher education levels (Graduate, Postgraduate, Master/Doctorate) show
minimal absenteeism. The Fisher's Exact Test produced a p-value of 0.8036, which
is much greater than 0.05.
This means there is no significant evidence to suggest that education levels and
absenteeism levels are correlated.
```

Observations for Question (e)

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

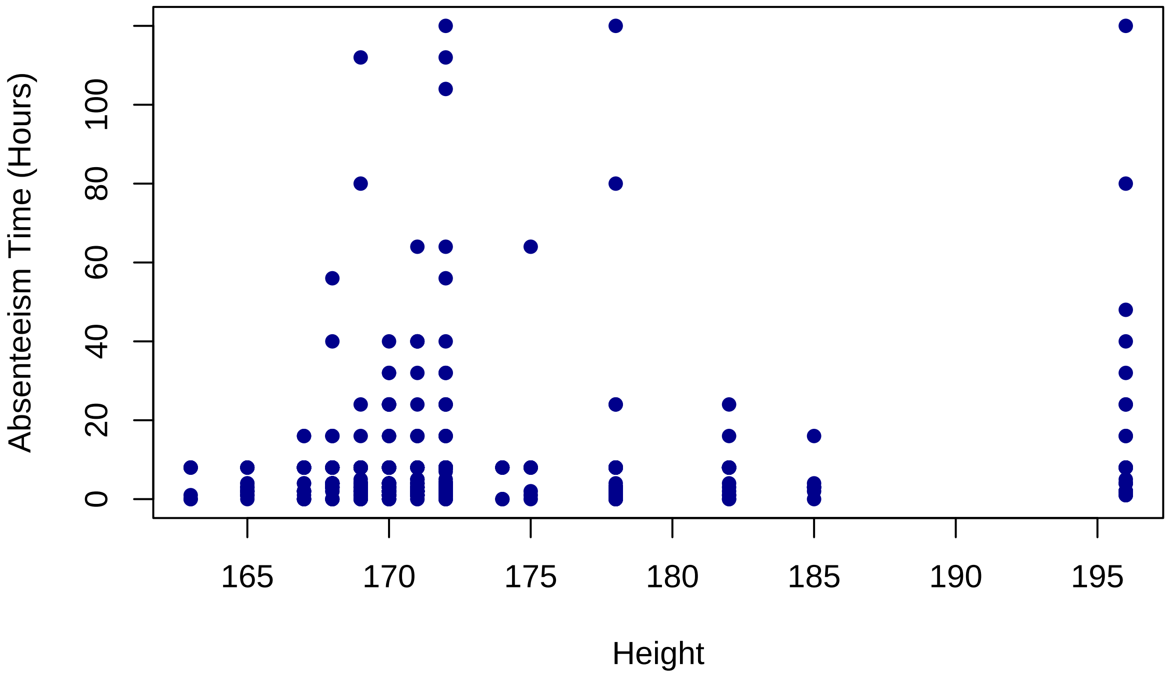
Explanation To discover which variables are most correlated with Absenteeism_time_in_hours, we compute Pearson correlation coefficients of absenteeism time with all other numeric variables in the data set. This will identify for us the variables most correlated with absenteeism. These can now be used as possible predictors in a logistic regression model.

```r
# Compute correlation coefficients for numeric variables
correlation_matrix <- cor(absent[, sapply(absent, is.numeric)], use = "complete.obs")

# Extract correlations with Absenteeism_time_in_hours
absenteeism_correlations <- correlation_matrix["Absenteeism_time_in_hours", ]

# Sort correlations in descending order
sorted_correlations <- sort(absenteeism_correlations, decreasing = TRUE)

# Display the sorted correlations
print(sorted_correlations)
```

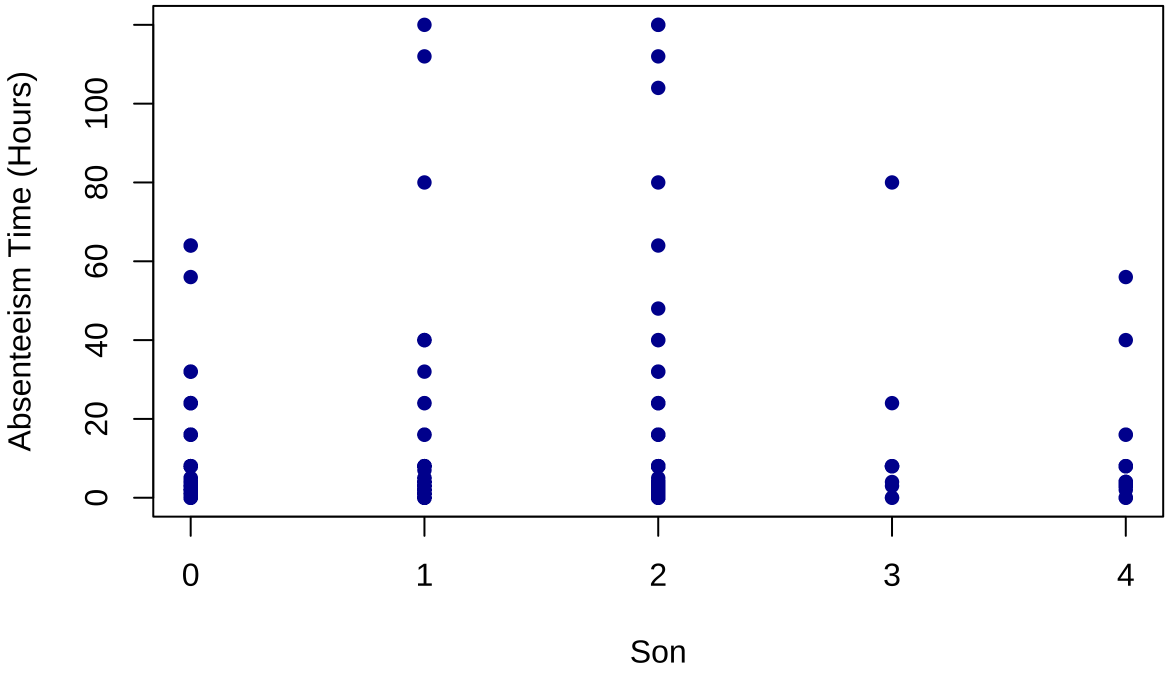
Code

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

```r
# Select top 5 correlated variables (excluding absenteeism itself)
top_variables <- names(sorted_correlations)[2:6]

# Create scatter plots for the top correlated variables using Base R
for (var in top_variables) {
  # Create the scatter plot
  plot(absent[[var]], absent$Absenteeism_time_in_hours,
       main = paste("Scatter Plot:", var, "vs Absenteeism Time"),
       xlab = var,
       ylab = "Absenteeism Time (Hours)",
       col = "darkblue",
       pch = 16) # Use circular points
}
```

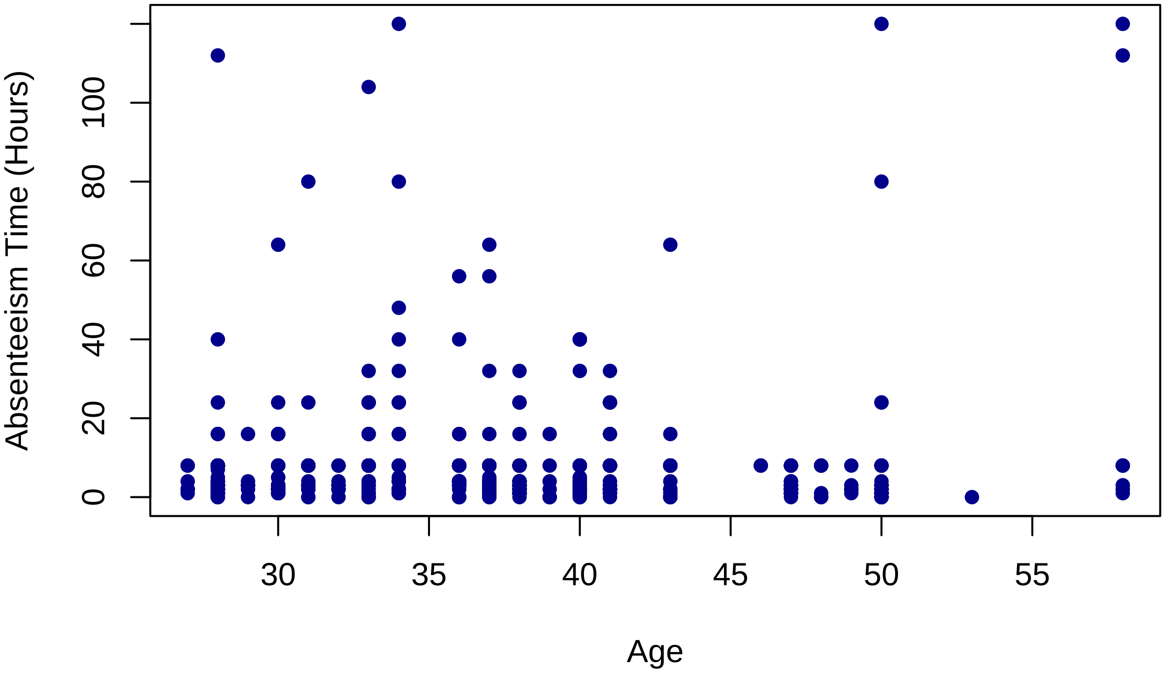


Scatter Plot: Height vs Absenteeism Time

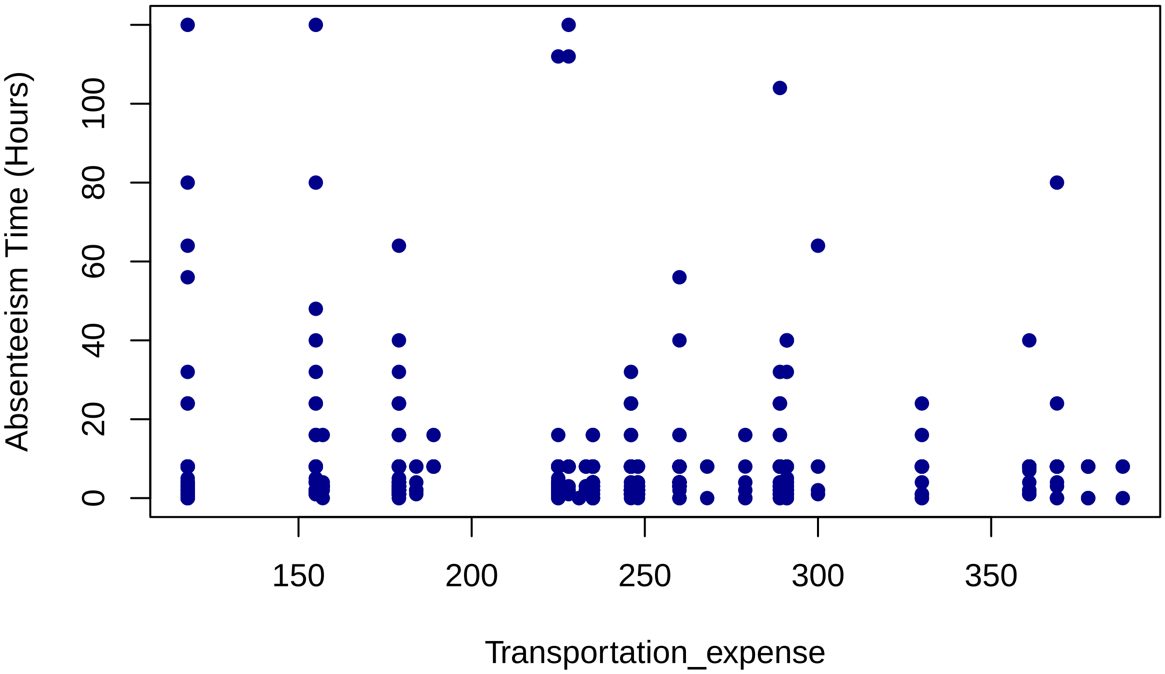


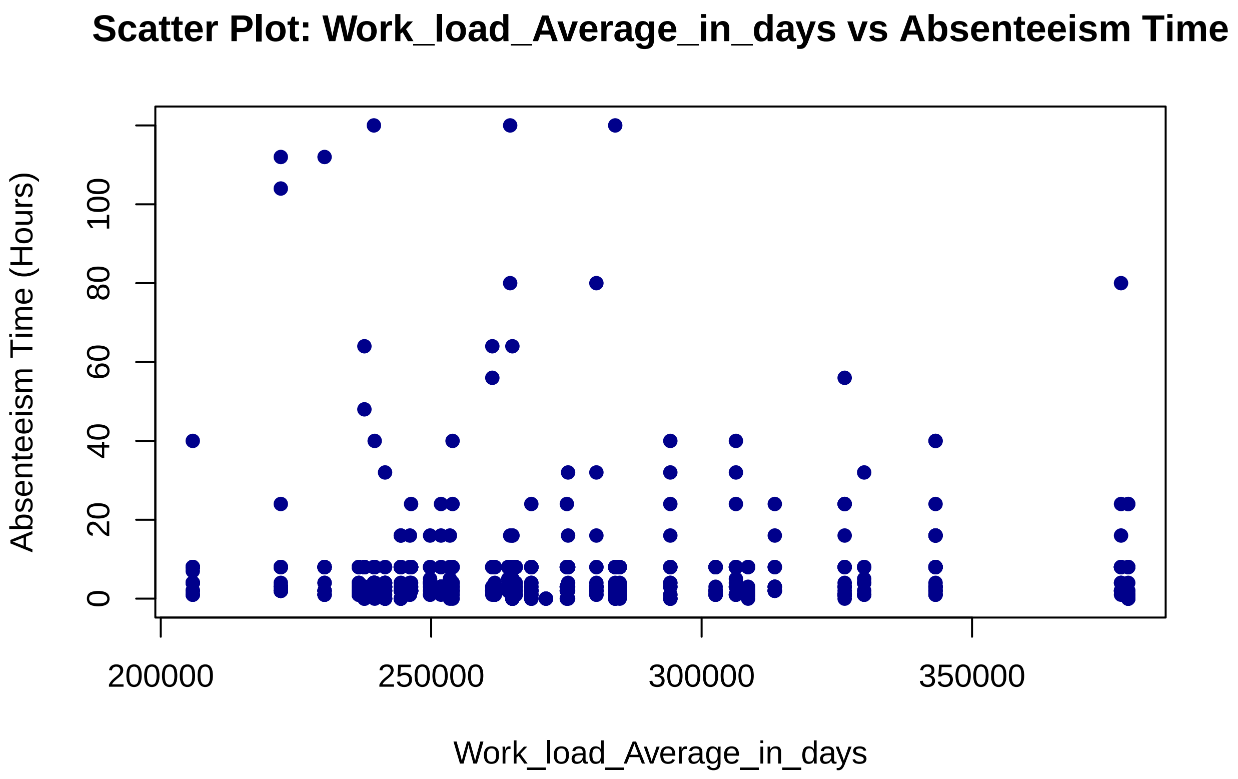
Scatter Plot: Son vs Absenteeism Time

Scatter Plot: Age vs Absenteeism Time



Scatter Plot: Transportation_expense vs Absenteeism Time





```
Height (0.144):
- Shows the strongest positive correlation among the variables. Taller individuals
seem to have slightly higher absenteeism, but the relationship is weak.

Son (0.114):
- Indicates a weak positive correlation; individuals with more sons might have
slightly higher absenteeism, but again, the effect is minimal.

Age (0.066):
Weak positive correlation; older individuals show slightly higher absenteeism,
but this trend is not strong.

Transportation_expense (0.028) and Work_load_Average_in_days (0.025):
Very weak positive correlations; these variables do not strongly impact absenteeism time.

Conclusion:
Most highly correlated with absenteeism time are Height, Son, and Age, though the
correlations are weak. These variables might contribute some predictive power to
a logistic regression model incorporating them, but their impact is small.
There could be other variables or interactions that play a more important role
in absenteeism.
```

Observations for Question (f)

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

```
Yes, analysis revealed some unexpected correlations:

Height:
Height correlated most strongly with absenteeism time (0.144), and this is surprising
since body height is not generally considered to be a factor in workplace absenteeism.

Son:
The number of sons had a weak positive correlation (0.114). This is contrary to
the assumption that family responsibilities (having children) are likely to
affect absenteeism in some predictable way.

Workload:
Workload-absenteeism relationship was not significant (0.025). This is contrary
to expectations because higher workloads are expected to lead to absenteeism due
to stress or burnout.

Transportation expenses:
Transportation cost showed a poor relationship (0.028), contrary to the expectation
that higher transportation costs lead to higher absenteeism.

Conclusion
These findings suggest that some variables commonly thought to influence absenteeism,
such as workload or travel cost, will have little effect on absenteeism. By contrast,
the height and number of sons correlation is anomalous and possibly subject to
further investigation in discounting coincidental or indirect influences.
```

Answer for Question (g)

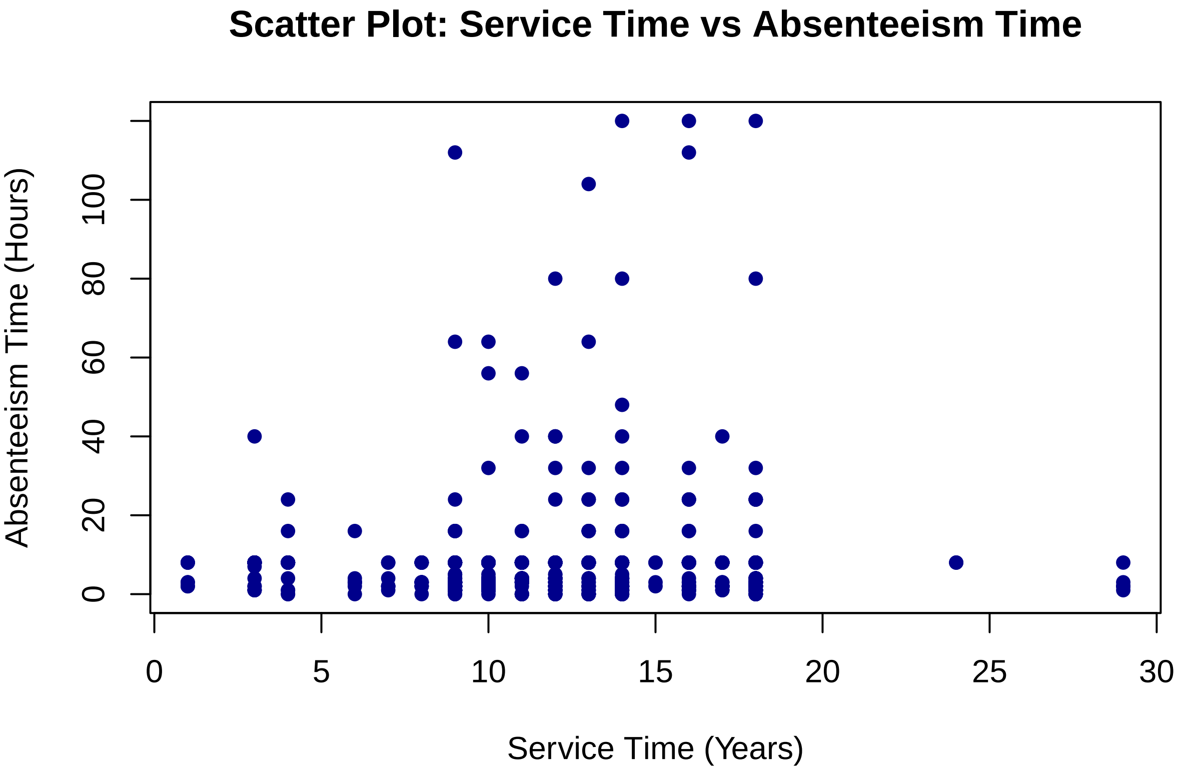
Question (h): Does service time (duration of service in the company) have any impact on the absenteeism rate?

```r
# Compute correlation coefficients for numeric variables
correlation_matrix <- cor(absent[, sapply(absent, is.numeric)], use = "complete.obs")

# Extract correlation between service time and absenteeism time
service_time_correlation <- correlation_matrix["Absenteeism_time_in_hours",
                                              "Service_time"]

# Print the correlation coefficient
print(paste("Correlation between Service Time and Absenteeism Time: ",
            round(service_time_correlation, 3)))
```

Code

```
## [1] "Correlation between Service Time and Absenteeism Time:  0.019"
```

```r
# Scatter plot for Service Time vs Absenteeism Time using Base R
plot(absent$Service_time, absent$Absenteeism_time_in_hours,
     main = "Scatter Plot: Service Time vs Absenteeism Time",
     xlab = "Service Time (Years)",
     ylab = "Absenteeism Time (Hours)",
     col = "darkblue",
     pch = 16) # Use circular points
```



```
Correlation Analysis:
The correlation between the time of service and the time of absenteeism in hours
was 0.019, and it is close to zero. This indicates that there is no linear relation
between the variables.

Scatter Plot Interpretation:
- The scatter plot confirms that there is no visible pattern or trend. The points
are comparatively scattered, suggesting no systematic relationship between the
time of service and the time of absenteeism.
- Staff within different durations of service tend to show equivalent behaviors
for absenteeism, implying that service time itself has minimal direct effects on
absenteeism.

Conclusion:
The implication of the findings is that the service time cannot play an instrumental
role in significantly affecting the absenteeism rate. This disredits any precepts
that employees working longer can or should have greater or lesser instances of
absenteeism than their counterworkers working short. Possibly more contributing
factors need to be factored into their decision.
```

Observations for Question (h)

Question (i): Examine if day of the week has any influence on absenteeism –

###. are certain days more prone to absences?

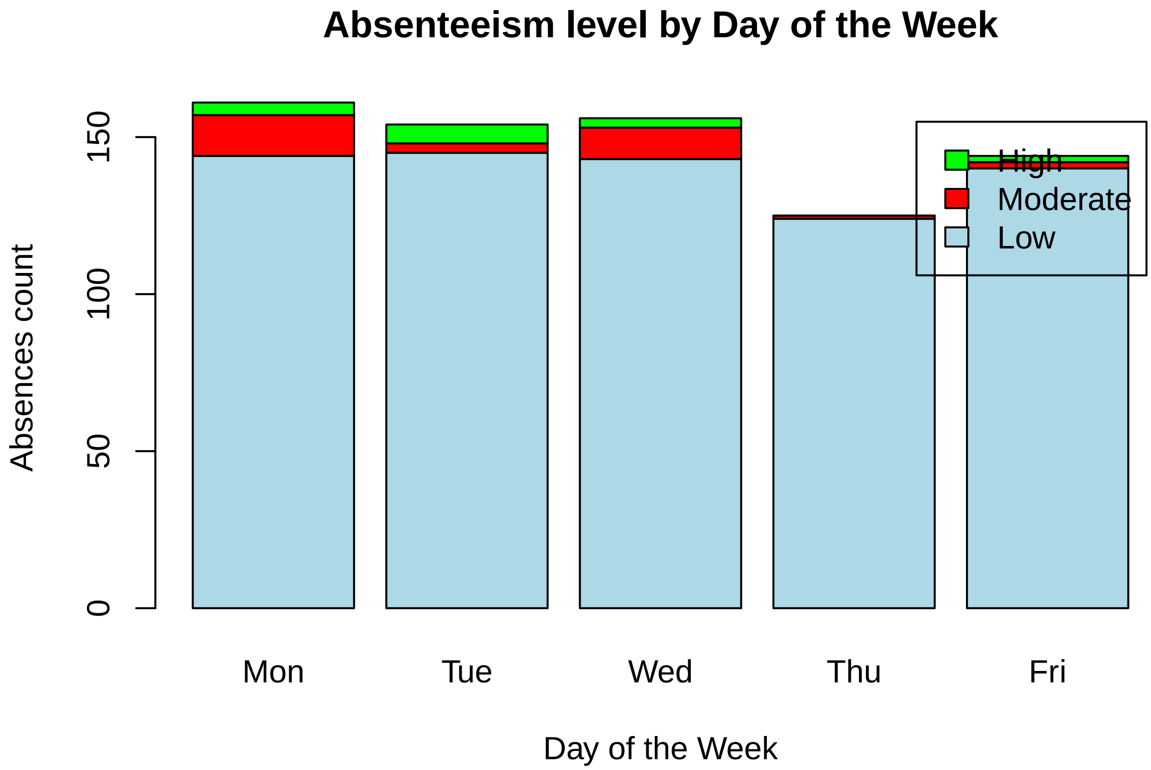
```
# Create a contingency table: absenteeism levels by day of the week
day_absenteeism <- table(absent$Day_of_the_week, absent$absenteeism)
```

```
print(day_absenteeism)
```

Code

```
##
##             Low Moderate High
##   Monday    144       13    4
##   Tuesday   145        3    6
##   Wednesday 143       10    3
##   Thursday  124        1    0
##   Friday    140        2    2
```

```
# Barplot to visualize absenteeism by day of the week
barplot(t(day_absenteeism),
        main = "Absenteeism level by Day of the Week",
        col = c("lightblue", "red", "green"),
        xlab = "Day of the Week",
        ylab = "Absences count",
        legend = TRUE,
        names.arg = c("Mon", "Tue", "Wed", "Thu", "Fri"))
```



```
# Perform Fisher's Exact Test with increased workspace
fisher_result <- fisher.test(day_absenteeism, workspace = 2e8) # Increase workspace
print(fisher_result)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  day_absenteeism
## p-value = 0.002095
## alternative hypothesis: two.sided
```

```r
# If the problem persists, use Monte Carlo simulation
fisher_result_simulated <- fisher.test(day_absenteeism, simulate.p.value = TRUE, B = 10000)
print(fisher_result_simulated)
```

```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  10000 replicates)
##
## data:  day_absenteeism
## p-value = 0.0021
## alternative hypothesis: two.sided
```

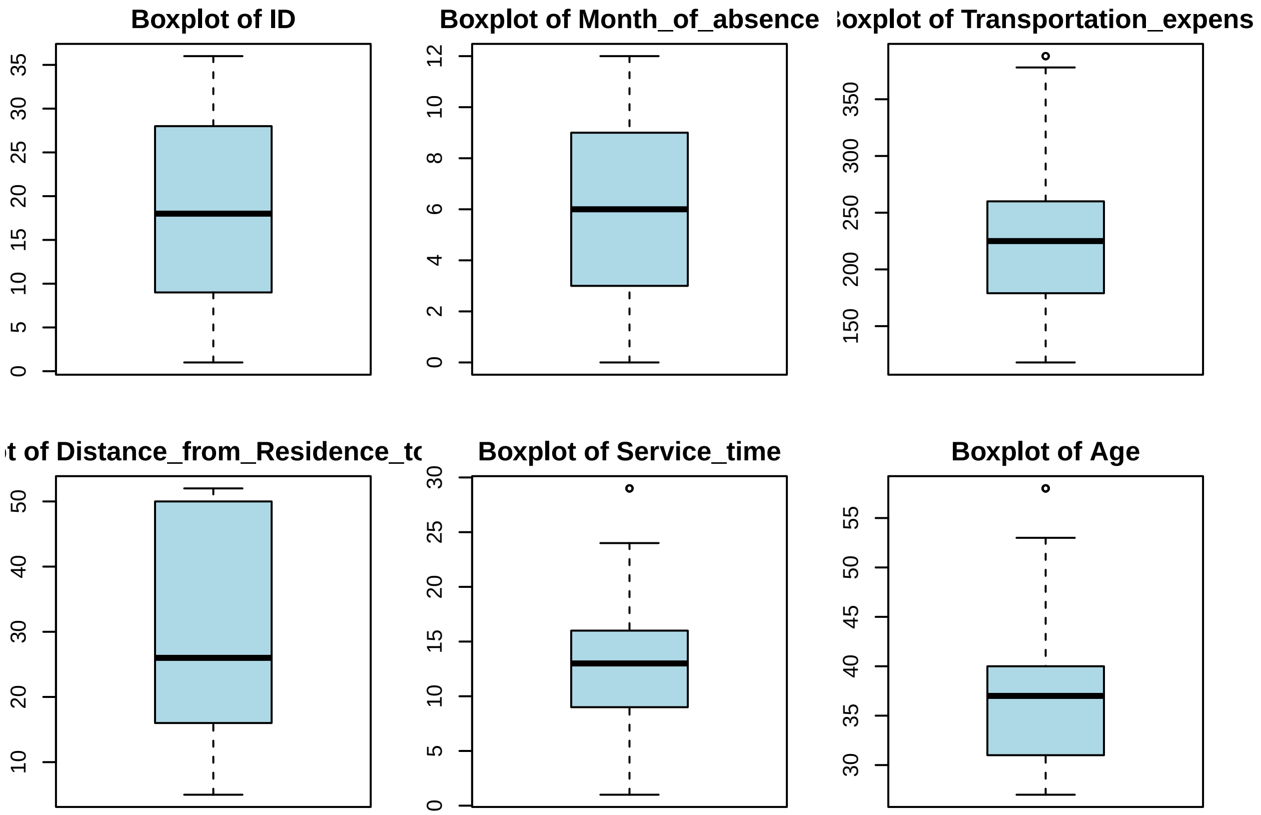
```

Copilot said: The Fisher's Exact Test (p = 0.0021)
The Fisher's Exact Test (p = 0.0021) and the simulated version (p = 0.0017) show
a significant association between the day of the week and absenteeism levels.
Absenteeism is higher on Monday, Tuesday, and Wednesday compared to Thursday and
Friday, especially in the "Low" category.
```

Observations for Question (i)

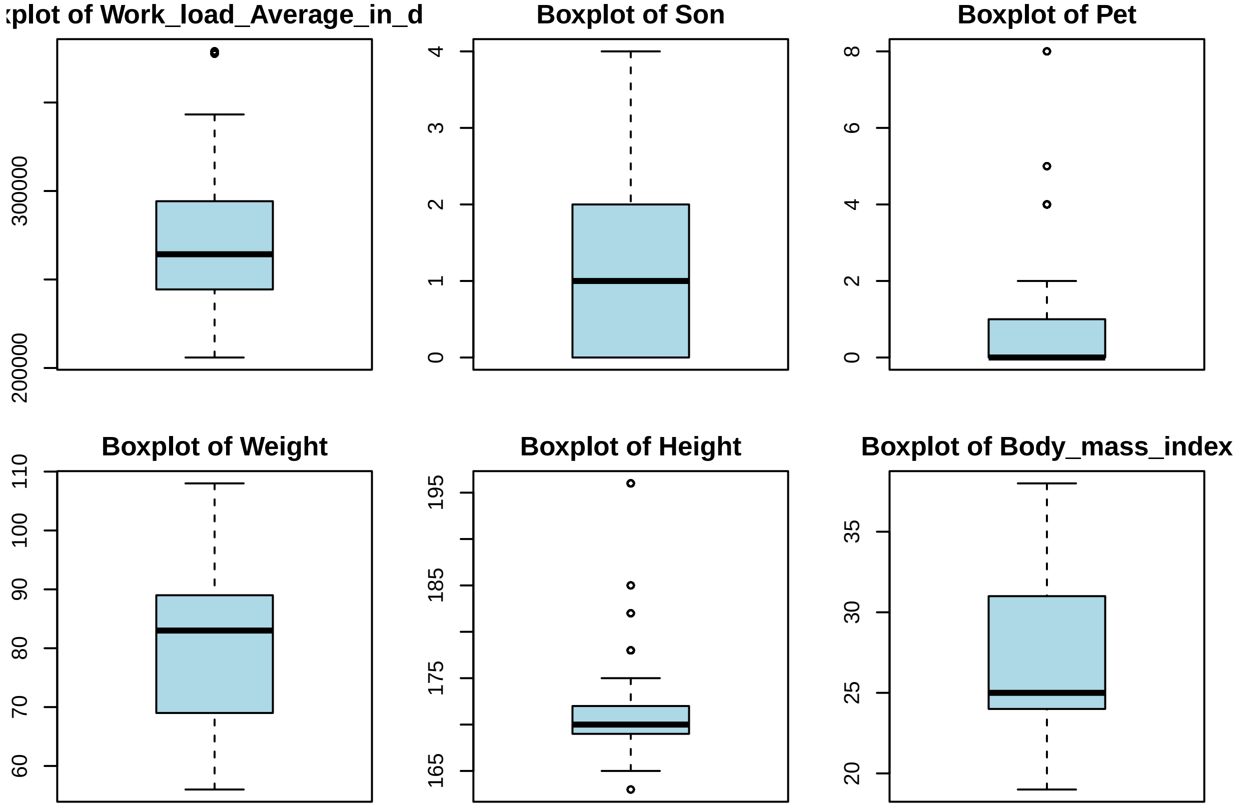
Question (j): Identify any outliers in the data set. What could be the reasons

###. for these anomalies, and how might they affect the analysis?

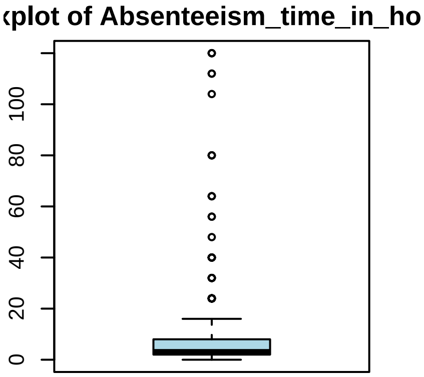
```r

# Extract numeric variables from the dataset
numeric_data <- absent[sapply(absent, is.numeric)]

# Adjust margins and layout for plotting
par(mar = c(2, 2, 2, 2))  # Set smaller margins
par(mfrow = c(2, 3))      # Set layout for fewer plots per frame

# Visualize numeric variables using boxplots
for (col_name in colnames(numeric_data)) {
  boxplot(numeric_data[[col_name]],
          main = paste("Boxplot of", col_name),
          col = "lightblue",
          outline = TRUE)
}
```



Code



```
# Reset layout
par(mfrow = c(1, 1))  # Reset to single plot layout
```



```r
# Function to detect outliers using the IQR rule
detect_outliers <- function(column_data) {
  Q1 <- quantile(column_data, 0.25, na.rm = TRUE)
  Q3 <- quantile(column_data, 0.75, na.rm = TRUE)
  IQR_value <- Q3 - Q1
  lower_limit <- Q1 - 1.5 * IQR_value
  upper_limit <- Q3 + 1.5 * IQR_value

  # Return indices of outliers
  return(which(column_data < lower_limit | column_data > upper_limit))
}

# Apply the outlier detection function to all numeric variables
outliers <- lapply(numeric_data, detect_outliers)

# Display results
print("Outliers in each numeric variable:")
```

```
## [1] "Outliers in each numeric variable:"
```

```r
print(outliers)
```

```
## $ID
## integer(0)
##
## $Month_of_absence
## integer(0)
##
## $Transportation_expense
## [1] 145 146 217
##
## $Distance_from_Residence_to_Work
## integer(0)
##
## $Service_time
## [1] 235 508 511 514 577
##
## $Age
## [1] 256 435 522 621 623 641 728 730
##
## $Work_load_Average_in_days
##  [1] 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223
## [20] 224 225 226 227 228 229 230 231 232 233 234 235 236
```

```
##
## $Son
## integer(0)
##
## $Pet
##  [1]   7  23  26  32  34  36  39  79 106 110 201 204 216 220 222 233 247 269 311
## [20] 323 337 398 400 431 470 475 507 541 555 562 565 595 604 630 657 664 691 702
## [39] 711 713 715 722 725 727 733 738
##
## $Weight
## integer(0)
##
## $Height
##   [1]   2   9  21  28  32  34  39  45  56  65  86  90 101 117 141 145 146 152
##  [19] 153 158 166 168 187 189 193 200 204 206 207 211 215 217 219 227 229 231
##  [37] 232 237 238 241 243 244 245 246 254 262 264 267 285 294 307 310 314 317
##  [55] 324 326 344 356 361 363 367 385 403 404 405 408 415 418 420 421 437 455
##  [73] 459 464 465 470 474 490 499 509 529 535 539 565 570 580 593 595 604 609
##  [91] 610 618 629 649 663 669 671 673 676 677 682 684 686 687 689 693 694 698
## [109] 699 703 705 712 713 715 716 717 722 725 727
##
## $Body_mass_index
## integer(0)
##
## $Absenteeism_time_in_hours
##  [1]   9  23  50  86  88  97 100 105 164 165 174 188 192 198 200 213 219 223 232
## [20] 250 273 280 288 300 324 360 388 421 448 449 472 535 556 570 571 623 649 653
## [39] 661 683 693 712 730 735
```

```
Transportation Expense: Rows [145, 146, 217]
Reason: High costs from relocation or private transport.
Impact: Inflates average costs.

Service Time: Rows [235, 508, 511, 514, 577]
Reason: Long tenure or data errors.
Impact: Distorts tenure trends.

Age: Rows [256, 435, ..., 730]
Reason: Very young or retired employees.
Impact: Skews age analysis.

Work Load (Average): Rows [205-236]
Reason: Temporary projects or errors.
Impact: Misrepresents workload patterns.

Pet: Extensive list of rows [7, ..., 738]
Reason: Data placeholders or reporting errors.
Impact: False correlations.

Height: Extensive list of rows [2, ..., 727]
Reason: Unit errors or extreme values.
Impact: Misleading health metrics.
```

```
Absenteeism Time: Rows [9, ..., 735]
Reason: Medical/parental leave or sabbaticals.
Impact: Overwhelms absenteeism models.

Recommendations:
-Validate outliers with domain experts.
-Remove or adjust anomalies as needed.
-Analyze separately for unique insights.
```

Observations for Question (j) #____ # Logistic Regression Analysis Instructions #____

(a) Building the Logistic Regression Model

Explanation: We are building a multinomial logistic regression model with the recoded absenteeism categories ("Low", "Moderate", "High") as the response variable. Categorical variables will be handled using appropriate encoding.

####Code:

```
# Build the model excluding ID and Absenteeism_time_in_hours
model <- multinom(absenteeism ~ . -ID -Absenteeism_time_in_hours, data = absent,
                  trace = FALSE)

# Summary of the model
summary(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
```

(b) Interpreting Coefficients of `son` and `weight`

Explanation:

- **`son`**: Represents the number of children an employee has.
- **`weight`**: Represents the weight of the employee.

####Code:

```
# Extract and display the coefficients for 'son' and 'weight'
coef_summary <- coef(model)  # Extract coefficients as a matrix

# Print the coefficients for 'son'
print("Coefficients for Son:")
```

```
## [1] "Coefficients for Son:"
```

```
print(exp(coef_summary[, "Son"]))
```

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

```
# Print the coefficients for 'weight'
print("Coefficients for Weight:")
```

```
## [1] "Coefficients for Weight:"
```

```
print(exp(coef_summary[, "Weight"]))
```

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

```
Son:
Moderate Absenteeism: Odds decrease by ~9.9% for each additional child (OR = 0.901).
High Absenteeism: Odds increase by ~107% for each additional child (OR = 2.070).

Weight:
Moderate Absenteeism: Odds increase by ~55.6% for each unit of weight (OR = 1.556).
High Absenteeism: Odds increase by ~78.7% for each unit of weight (OR = 1.787).


More children lead to more chances of having "High" absenteeism for family reasons.
Employees of higher weight tend to have "Moderate" or "High" absenteeism, possibly
as a result of health issues.
```

Observations for Question (b)

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

####Code:

```
# Perform backward selection using stepAIC
final_model <- stepAIC(model, direction = "backward", trace = TRUE)
```

```
## Start:  AIC=394.55
## absenteeism ~ (ID + Month_of_absence + Day_of_the_week + Seasons +
##     Transportation_expense + Distance_from_Residence_to_Work +
##     Service_time + Age + Work_load_Average_in_days + Education +
##     Son + Pet + Weight + Height + Body_mass_index + Absenteeism_time_in_hours) -
##     ID - Absenteeism_time_in_hours
##
##                                   Df    AIC
## - Seasons                          6 387.91
## - Education                        6 388.21
## - Service_time                     2 392.46
## - Age                              2 392.61
## - Transportation_expense           2 393.63
## - Month_of_absence                 2 393.66
## - Height                           2 393.90
## <none>                               394.55
## - Son                              2 395.02
## - Weight                           2 395.03
## - Distance_from_Residence_to_Work  2 395.03
## - Work_load_Average_in_days        2 395.19
## - Body_mass_index                  2 395.68
## - Pet                              2 396.07
## - Day_of_the_week                  8 403.58
##
## Step:  AIC=387.91
## absenteeism ~ Month_of_absence + Day_of_the_week + Transportation_expense +
##     Distance_from_Residence_to_Work + Service_time + Age + Work_load_Average_in_days +
##     Education + Son + Pet + Weight + Height + Body_mass_index
##
##                                   Df    AIC
## - Education                        6 380.98
## - Age                              2 385.37
## - Month_of_absence                 2 385.41
## - Service_time                     2 385.44
## - Transportation_expense           2 386.93
## - Height                           2 387.07
## <none>                               387.91
## - Son                              2 387.95
## - Distance_from_Residence_to_Work  2 388.07
## - Weight                           2 388.20
## - Body_mass_index                  2 388.81
## - Pet                              2 389.13
## - Work_load_Average_in_days        2 389.69
## - Day_of_the_week                  8 396.50
##
## Step:  AIC=380.98
## absenteeism ~ Month_of_absence + Day_of_the_week + Transportation_expense +
##     Distance_from_Residence_to_Work + Service_time + Age + Work_load_Average_in_days +
##     Son + Pet + Weight + Height + Body_mass_index
##
##                                   Df    AIC
## - Height                           2 378.68
## - Service_time                     2 378.81
## - Month_of_absence                 2 378.94
## - Age                              2 379.41
## - Weight                           2 379.80
## - Body_mass_index                  2 380.23
## - Distance_from_Residence_to_Work  2 380.79
## <none>                               380.98
## - Transportation_expense           2 381.02
## - Pet                              2 381.76
## - Son                              2 381.99
## - Work_load_Average_in_days        2 383.71
## - Day_of_the_week                  8 389.79
##
## Step:  AIC=378.68
## absenteeism ~ Month_of_absence + Day_of_the_week + Transportation_expense +
##     Distance_from_Residence_to_Work + Service_time + Age + Work_load_Average_in_days +
##     Son + Pet + Weight + Body_mass_index
##
##                                   Df    AIC
## - Service_time                     2 376.58
## - Month_of_absence                 2 376.83
## - Age                              2 377.04
## - Distance_from_Residence_to_Work  2 378.02
## - Pet                              2 378.55
## <none>                               378.68
## - Son                              2 379.18
## - Transportation_expense           2 379.29
## - Weight                           2 381.13
## - Work_load_Average_in_days        2 381.95
## - Body_mass_index                  2 383.00
## - Day_of_the_week                  8 386.53
##
```

```
## Step:  AIC=376.58
## absenteeism ~ Month_of_absence + Day_of_the_week + Transportation_expense + 
##     Distance_from_Residence_to_Work + Age + Work_load_Average_in_days + 
##     Son + Pet + Weight + Body_mass_index
## 
##                                   Df    AIC
## - Month_of_absence                 2 374.93
## - Age                              2 375.15
## - Transportation_expense           2 375.92
## - Son                              2 376.28
## <none>                               376.58
## - Distance_from_Residence_to_Work  2 377.23
## - Pet                              2 377.83
## - Weight                           2 379.47
## - Work_load_Average_in_days        2 380.61
## - Body_mass_index                  2 380.75
## - Day_of_the_week                  8 384.76
## 
## Step:  AIC=374.93
## absenteeism ~ Day_of_the_week + Transportation_expense + Distance_from_Residence_to_Work + 
##     Age + Work_load_Average_in_days + Son + Pet + Weight + Body_mass_index
## 
##                                   Df    AIC
## - Age                              2 373.15
## - Transportation_expense           2 374.46
## - Son                              2 374.72
## <none>                               374.93
## - Distance_from_Residence_to_Work  2 375.53
## - Pet                              2 375.96
## - Weight                           2 377.58
## - Body_mass_index                  2 378.47
## - Work_load_Average_in_days        2 378.69
## - Day_of_the_week                  8 383.04
## 
## Step:  AIC=373.15
## absenteeism ~ Day_of_the_week + Transportation_expense + Distance_from_Residence_to_Work + 
##     Work_load_Average_in_days + Son + Pet + Weight + Body_mass_index
## 
##                                   Df    AIC
## - Transportation_expense           2 373.05
## <none>                               373.15
## - Weight                           2 374.42
## - Pet                              2 374.75
## - Body_mass_index                  2 374.76
## - Son                              2 375.39
## - Distance_from_Residence_to_Work  2 375.44
## - Work_load_Average_in_days        2 378.30
## - Day_of_the_week                  8 381.15
## 
## Step:  AIC=373.05
## absenteeism ~ Day_of_the_week + Distance_from_Residence_to_Work + 
##     Work_load_Average_in_days + Son + Pet + Weight + Body_mass_index
## 
##                                   Df    AIC
## - Pet                              2 372.33
## <none>                               373.05
## - Weight                           2 373.71
## - Body_mass_index                  2 374.68
## - Son                              2 374.82
## - Distance_from_Residence_to_Work  2 375.69
## - Work_load_Average_in_days        2 378.50
## - Day_of_the_week                  8 381.30
##
## Step:  AIC=372.33
## absenteeism ~ Day_of_the_week + Distance_from_Residence_to_Work +
##     Work_load_Average_in_days + Son + Weight + Body_mass_index
##
##                                   Df    AIC
## <none>                               372.33
## - Weight                           2 373.48
## - Son                              2 373.94
## - Body_mass_index                  2 374.16
## - Distance_from_Residence_to_Work  2 374.80
## - Work_load_Average_in_days        2 377.63
## - Day_of_the_week                  8 379.91
```

```
# Summary of the final selected model
summary(final_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
```

```
# Extract the predictors retained in the final model
final_predictors <- names(coef(final_model))
print("Predictors retained in the final model:")
```

```
## [1] "Predictors retained in the final model:"
```

```
print(final_predictors)
```

```
## NULL
```

```

The final model keeps variables that are significant predictors of absenteeism
with minimal model complexity (AIC = 372.33). The predictors Seasons, Education,
and Height were excluded from the selection.

The predictors kept indicate that absenteeism is influenced by:
- Day of the Week: Specific days might have absenteeism patterns that are higher.
- Distance to Work: Longer distances might influence the level of absenteeism.
- Workload: Average workload might influence absenteeism.
- Number of Children (Son): Employees with more children will be more prone to
absenteeism.
- Weight and BMI: Medical reasons could be a cause of absenteeism.
```

Observations for Question (c)

(d) Interpret some (at least 2) of your model's findings in a practical workplace context.(Formulate recommendations for an organization based on these findings.

```

*******Finding 1: Day_of_the_week*******

The Day_of_the_week predictor has an impact on absenteeism.
Employees tend to be absent on certain days (e.g., Mondays or Fridays) as
opposed to others, which points to a weekend-related absenteeism pattern.

Practical Recommendation:
- Flexible Work Schedules: Adopt flexible scheduling or remote work on Mondays
and Fridays to reduce absenteeism.
This might reduce "long weekend" absenteeism by providing employees with more
autonomy over their work schedule.
- Engagement Strategies: Increase attendance on such days through team-building
exercises or lightened workloads.

*******Finding 2: Distance_from_Residence_to_Work*******

Absenteeism is higher in those employees having longer travel times.
Physical fatigue, inability to travel, or time limitations may be the cause.
```

```
Practical Recommendation:
- Transportation Support: Offer transport support, i.e., shuttle travel, or travel
subsidy on public transport so that travel stress can be reduced.
- Hybrid or Remote Work: Facilitate telecommuting for a couple of days a week for
individuals with lengthy commute times, reducing the need to travel daily and
improving attendance.
```

Observations for Question (d) #________________________________ # Flu Shot Data Analysis #________________________________

Question (a): Create a scatterplot matrix of the data. What are your

###. observations?

Explanation We need to create a scatterplot matrix to visually examine the relationships among the variables:

Y: Flu shot status (1 = Received, 0 = Not Received) X1: Age X2: Health awareness index X3: Gender (1 = Male, 0 = Female)

This will help us observe potential correlations or patterns between predictors and the response variable.

```
flu = read.table("flu shot.txt", header = TRUE)

# Convert categorical variables to factors
flu$flu_shot = factor(flu$flu_shot, levels = c(0, 1), labels = c("No", "Yes"))
flu$sex = factor(flu$sex, levels = c(0, 1), labels = c("Female", "Male"))

# Explore structure
str(flu)
```

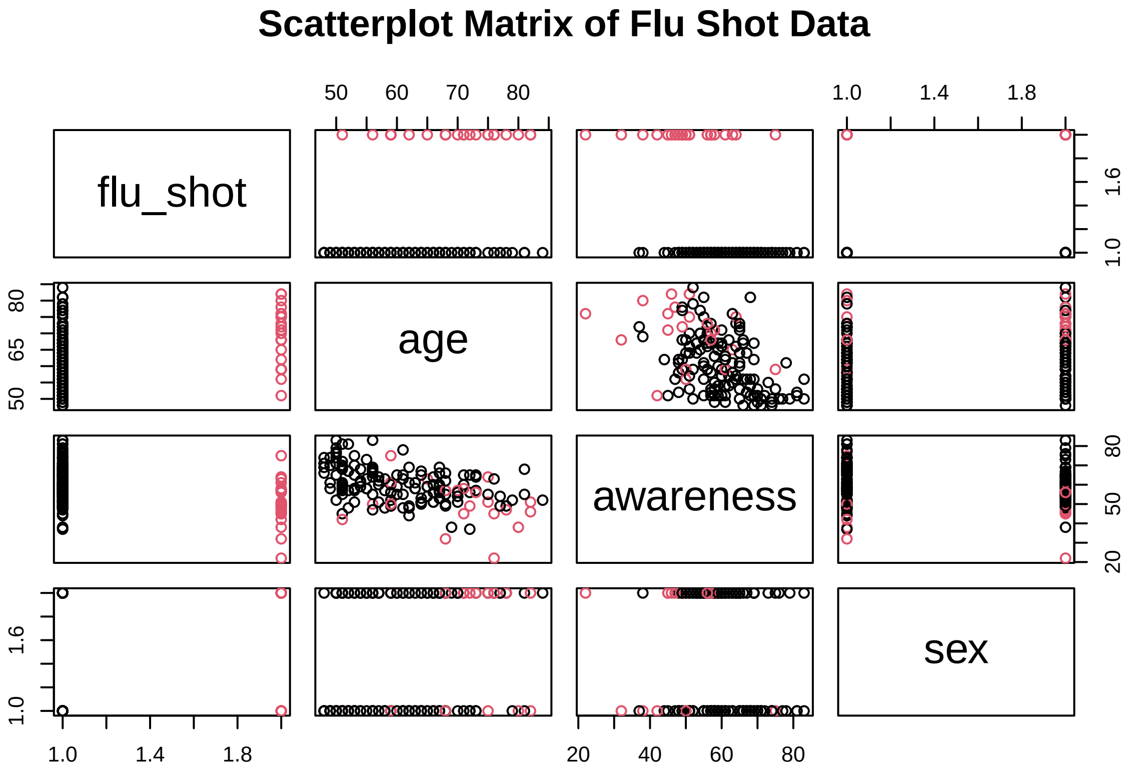
Code

```
## 'data.frame':    159 obs. of  4 variables:
##  $ flu_shot : Factor w/ 2 levels "No","Yes": 1 1 2 1 1 1 1 1 1 1 ...
##  $ age      : int  59 61 82 51 53 62 51 70 71 55 ...
##  $ awareness: int  52 55 51 70 70 49 69 54 65 58 ...
##  $ sex      : Factor w/ 2 levels "Female","Male": 1 2 1 1 1 2 2 2 2 2 ...
```

```
# Logistic regression model
flu_model = glm(flu_shot ~ age + awareness + sex, data = flu, family = binomial)
summary(flu_model)
```

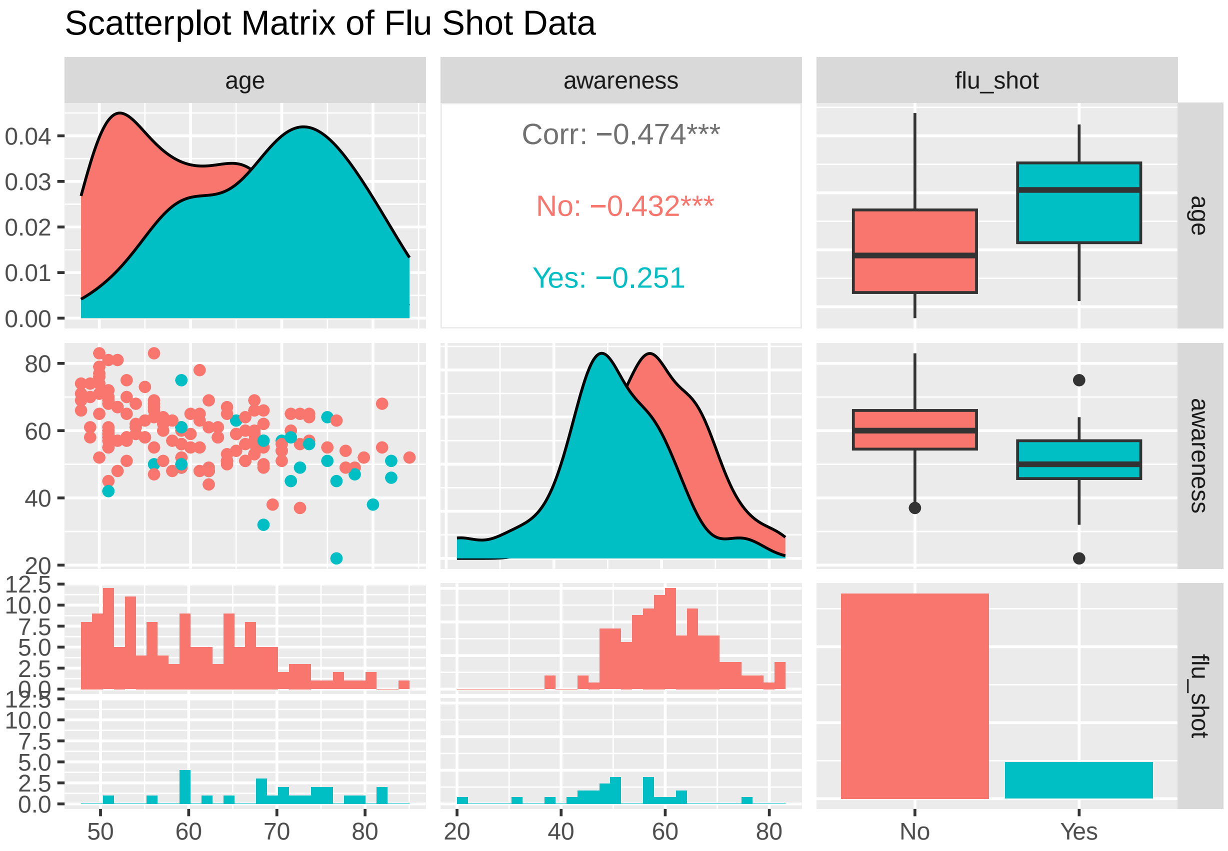
```
##
## Call:
## glm(formula = flu_shot ~ age + awareness + sex, family = binomial,
##     data = flu)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.17716    2.98242  -0.395  0.69307
## age          0.07279    0.03038   2.396  0.01658 *
## awareness   -0.09899    0.03348  -2.957  0.00311 **
## sexMale      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
```

```
# Create a scatterplot matrix including both numeric and categorical variables
pairs(flu, main = "Scatterplot Matrix of Flu Shot Data", col = flu$flu_shot)
```



```
# Create scatterplot matrix
ggpairs(
  flu[, c("age", "awareness", "flu_shot")],
  aes(color = flu$flu_shot),
  title = "Scatterplot Matrix of Flu Shot Data"
)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
Scatterplot Matrix:
- Age vs. Awareness: A moderate negative correlation (-0.474), where older
individuals tend to have lower awareness.
- Flu Shot:
Older individuals and those with slightly higher awareness are more likely to
receive a flu shot.
- Gender: No significant difference in age or awareness between males and females.

Logistic Regression:
- Age: Positive predictor-older individuals are more likely to get a flu shot.
- Awareness: Negative predictor-higher awareness reduces the likelihood of getting
a flu shot.
- Gender: Not a significant factor.

Conclusion:
Age and awareness significantly influence flu shot uptake, while gender does not.
```

Observations for Question (a)

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

Explanation Multiple logistic regression models the relationship between a binary dependent variable and multiple independent variables. It helps to understand how each predictor influences the probability of an event occurring.

```r
# Fit the logistic regression model
flu_model = glm(flu_shot ~ age + awareness + sex, data = flu, family = binomial)

# Display the summary of the model
summary(flu_model)
```

Code

```
##
## Call:
## glm(formula = flu_shot ~ age + awareness + sex, family = binomial,
##     data = flu)
##
## 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 **
## sexMale      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
```

```
Intercept:
Estimate: -1.17716
-The intercept represents the baseline log-odds of receiving a flu shot when all
predictors (age, awareness, and sex) are at their reference levels (e.g., age = 0,
awareness = 0, and sex = Female).

Predictor: (age)
Estimate: 0.07279
p-value: 0.01658 (Significant at the 5% level).
Interpretation: For each additional year of age, the log-odds of receiving a flu
shot increase by 0.07279. Older individuals are more likely to get vaccinated.

Predictor: (awareness)
Estimate: -0.09899
p-value: 0.00311 (Highly significant at the 1% level).
Interpretation: For each unit increase in the awareness index, the log-odds of
receiving a flu shot decrease by 0.09899. Higher awareness is associated with a
lower likelihood of getting vaccinated.

Predictor: (sex)
Estimate: 0.43397
p-value: 0.40558 (Not significant).
```

```
Interpretation: The gender of the individual (Male vs. Female) does not
significantly impact the likelihood of receiving a flu shot.
```

Observation of Question (b)

Question (c): State the fitted regression equation.

Explanation The fitted regression equation represents the relationship between the dependent variable and the independent variables in a logistic regression model.

```
# Fitted regression equation
beta <- coef(flu_model)
cat("Fitted regression equation: logit(flu_shot) =", round(beta[1], 2), "+",
    round(beta[2], 2), "* age +", round(beta[3], 2), "* awareness +",
    round(beta[4], 2), "* sex")
```

Code

```
## Fitted regression equation: logit(flu_shot) = -1.18 + 0.07 * age + -0.1 * awareness + 0.43 * sex
```

```
Intercept: When age = 0, awareness = 0, and sex = Female, the log-odds of
receiving a flu shot are -1.18.

Age: For every 1-year increase in age, the log-odds of receiving a flu shot
increase by 0.07. Older individuals are more likely to get vaccinated.

Awareness: For every 1-unit increase in the awareness index, the log-odds of
receiving a flu shot decrease by 0.1. Higher awareness is associated with a
lower likelihood of getting vaccinated.

Sex (Male): Males have 0.43 higher log-odds of receiving a flu shot compared to
females.
```

Observations for Question (c)

Question (d): Obtain exp(1), exp(2), exp(3) and interpret these numbers.

Explanation The exponentiated coefficients (exp(beta)) represent the odds ratios, which indicate how the odds of the dependent variable change with a one-unit increase in the predictor variable.

```
# Exponentiated coefficients
exp_beta <- exp(coef(flu_model))
exp_beta
```

Code

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

```
Age (exp( 1)=1.0755):
For every 1-year increase in age, the odds of receiving a flu shot increase by
approximately 7.55%, holding all other variables constant.
```

```
Awareness (exp( 2)=0.9058):
For every 1-unit increase in the awareness index, the odds of receiving a flu
shot decrease by approximately 9.42% (1-0.9058),holding all other variables constant.

Sex (Male) (exp( 3)=1.5435):
Males have 1.54 times the odds (or are 54.35% more likely) of receiving a flu
shot compared to females, holding all other variables constant.
```

Observations for Question (d)

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

Explanation Predicting the probability of an event occurring based on the logistic regression model involves using the fitted coefficients and the values of the predictors.

```r
# Logistic regression model coefficients
beta <- coef(flu_model)

# Calculate odds ratios
odds_ratios <- exp(beta)
cat("Odds Ratios:\n")
```

Code

```
## Odds Ratios:
```

```r
print(odds_ratios)
```

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

```r
# Estimating the probability for a male client, age 55, awareness index 60
age <- 55
awareness <- 60
sex <- 1  # Male

logit <- beta[1] + beta[2] * age + beta[3] * awareness + beta[4] * sex
probability <- exp(logit) / (1 + exp(logit))

cat("\nEstimated Probability for male (age 55, awareness 60):", round(probability, 4), "\n")
```

```
##
## Estimated Probability for male (age 55, awareness 60): 0.0642
```

```
For a male patient aged 55 years and possessing a health awareness index score
of 60, the probability of being vaccinated for the flu shot is approximately
6.42%. This shows that in this situation, probabilities for vaccination are
still low.
```

Observations for Question (e)

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

Explanation The Wald test assesses the significance of individual predictors in the logistic regression model. It helps to determine if a predictor can be dropped from the model.

```
# Extract the coefficient and standard error for `sexMale`
coef_sex <- coef(flu_model)["sexMale"]
se_sex <- summary(flu_model)$coefficients["sexMale", "Std. Error"]

# Calculate the Wald test statistic (Z-value)
wald_statistic <- coef_sex / se_sex

# Calculate the p-value for the Wald test
p_value <- 2 * (1 - pnorm(abs(wald_statistic)))

# Print the results
cat("Wald Test Statistic for 'sexMale':", wald_statistic, "\n")
```

Code

```
## Wald Test Statistic for 'sexMale': 0.8316976
```

```
cat("P-value for Wald Test:", p_value, "\n")
```

```
## P-value for Wald Test: 0.4055797
```

```
# Check significance
alpha <- 0.05
if (p_value > alpha) {
  cat("The p-value is greater than", alpha, ". We fail to reject the null hypothesis.
      'sex' can be dropped from the model.\n")
} else {
  cat("The p-value is less than", alpha, ". We reject the null hypothesis. 'sex'
      should be kept in the model.\n")
}
```

```
## The p-value is greater than 0.05 . We fail to reject the null hypothesis.
##       'sex' can be dropped from the model.
```

```
The Wald test was conducted to determine if the gender variable is statistically
significant in predicting the likelihood of receiving a flu shot.

Wald Test Statistic: 0.8316976
P-value: 0.4055797

Conclusion:
The p-value is greater than 0.05 . We fail to reject the null hypothesis. 'sex'
can be dropped from the model.
```

Observations for Question (f)

Question (g): Use forward selection to decide which predictor variables enter should be kept in the regression model.Forward Selection for Predictor Variables

Explanation Forward selection is a stepwise regression method that starts with an empty model and adds significant predictors one by one. It helps to identify the most important predictors for the model.

```r
# Define the full model (all predictors)
flu_model <- glm(flu_shot ~ age + awareness + sex, family = binomial, data = flu)

# Perform forward selection using stepAIC
forward_model <- step(glm(flu_shot ~ 1, family = binomial, data = flu),
                      scope = list(lower = ~1, upper = ~age + awareness + sex),
                      direction = "forward")  # Forward selection
```

Code

```
## 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
```

```r
# Summary of the selected model
summary(forward_model)
```

```
##
## Call:
## glm(formula = flu_shot ~ awareness + age, family = binomial,
##     data = flu)
##
## 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
```

```
Awareness: Negative relationship ((p = 0.003)), reducing odds of flu shots as
awareness increases.
Age: Positive relationship ((p = 0.009)), increasing odds of flu shots with age.

Conclusion:
The final model includes awareness and age as predictors. Sex is not significant
and is excluded.
```

Observations for Question (g)

Question (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)?

Explanation Backward selection is a stepwise regression method that starts with a full model and removes non-significant predictors one by one. It helps to simplify the model by retaining only significant predictors.

```
# Define the full model (all predictors)
flu_model <- glm(flu_shot ~ age + awareness + sex, family = binomial, data = flu)

# Perform backward selection using stepAIC
backward_model <- step(flu_model, direction = "backward")
```

Code

```
## 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
```

```r
# Summary of the selected model
summary(backward_model)
```

```
##
## Call:
## glm(formula = flu_shot ~ age + awareness, family = binomial,
##     data = flu)
##
## 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
```

```
Comparison of Backward Selection Results to Part (f) (Wald Test for Gender):

Wald Test Result (Part f):
The Wald test showed that the gender variable ((X_3)) is not statistically
significant ((p = 0.4056 > 0.05)).

Conclusion: Gender can be dropped from the regression model.

Backward Selection Result (Part h):
In backward selection, gender ((X_3)) was also removed because it did not
significantly improve the model (AIC decreased when gender was excluded).

Final Model: Included age and awareness as significant predictors, excluding gender.

Comparison:
Both the Wald test and backward selection agree that gender ((X_3)) is not a
meaningful predictor and can be excluded from the regression model.
This consistency reinforces the conclusion that gender does not contribute
significantly to predicting the outcome.
```

Observations for Question (h)

Question (i): How would you interpret

ˆ 0, ˆ 1 and ˆ 3

Explanation Interpreting the coefficients in a logistic regression model helps to understand the impact of each predictor on the dependent variable. The coefficients represent the change in log odds for a one-unit increase in the predictor.

```r
# Fit logistic regression model
flu_model <- glm(flu_shot ~ age + awareness + sex, data = flu, family = binomial)

# Summary of the model
summary(flu_model)
```

Code

```
##
## Call:
## glm(formula = flu_shot ~ age + awareness + sex, family = binomial,
##     data = flu)
##
## 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 **
## sexMale      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
```

```r
# Extract coefficients
coefficients <- coef(flu_model)

# Interpret coefficients
beta_0 <- coefficients[1]  # Intercept
beta_1 <- coefficients[2]  # Age
beta_3 <- coefficients[4]  # Sex

# Convert to odds ratio for interpretation
exp_beta_0 <- exp(beta_0)  # Baseline odds when all predictors are 0
exp_beta_1 <- exp(beta_1)  # Effect of 1-unit increase in age on odds
exp_beta_3 <- exp(beta_3)  # Effect of gender (1 vs. 0) on odds

# Print interpretations
cat("Interpretation of coefficients:\n")
```

```
## Interpretation of coefficients:
```

```r
cat("1. exp(beta_0): Baseline odds =", exp_beta_0, "\n")
```

```
## 1. exp(beta_0): Baseline odds = 0.3081529
```

```r
cat("2. exp(beta_1): Odds ratio for 1-unit increase in age =", exp_beta_1, "\n")
```

```
## 2. exp(beta_1): Odds ratio for 1-unit increase in age = 1.075503
```

```r
cat("3. exp(beta_3): Odds ratio for sex (Male vs Female) =", exp_beta_3, "\n")
```

```
## 3. exp(beta_3): Odds ratio for sex (Male vs Female) = 1.54338
```