STAT 561 - Homework 2

Group Members: [List all group members here]

Due: April 14, 2025

Introduction

In this homework, we tackle the challenge of absenteeism in today's dynamic work environment using logistic regression. We aim to predict and understand the patterns of absenteeism using a detailed dataset from a courier company in Brazil. Additionally, we analyze flu shot data to understand factors influencing flu vaccination uptake.

Setup

Absenteeism Data Analysis

Question 1: Exploratory Data Analysis (EDA)

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 the dataset
absent <- read_excel("Absenteeism_at_work.xls")

# Recode absenteeism
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)
absent$Month_of_absence <- factor(absent$Month_of_absence)
absent$Day_of_the_week <- factor(absent$Day_of_the_week)
absent$Seasons <- factor(absent$Seasons)
absent$Education <- factor(absent$Education)


# Check for missing values
missing_values <- sapply(absent, function(x) sum(is.na(x)))
print(missing_values)
```

Code

```
##                             ID               Month_of_absence
```

```
##                              0                              0
##                Day_of_the_week                        Seasons
##                              0                              0
##          Transportation_expense Distance_from_Residence_to_Work
##                              0                              0
##                   Service_time                            Age
##                              0                              0
##       Work_load_Average_in_days                      Education
##                              0                              0
##                            Son                            Pet
##                              0                              0
##                         Weight                         Height
##                              0                              0
##                Body_mass_index       Absenteeism_time_in_hours
##                              0                              0
##                    absenteeism
##                              0
```

```
# View the structure of the data
str(absent)
```

```
## tibble [740 x 17] (S3: tbl_df/tbl/data.frame)
##  $ ID                              : num [1:740] 11 36 3 7 11 3 10 20 14 1 ...
##  $ Month_of_absence                : Factor w/ 13 levels "0","1","2","3",..: 8 8 8 8 8 8 8 8 8 8 ...
##  $ Day_of_the_week                 : Factor w/ 5 levels "2","3","4","5",..: 2 2 3 4 4 5 5 5 1 1 ...
##  $ Seasons                         : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 1 ...
##  $ Transportation_expense          : num [1:740] 289 118 179 279 289 179 361 260 155 235 ...
##  $ Distance_from_Residence_to_Work: num [1:740] 36 13 51 5 36 51 52 50 12 11 ...
##  $ Service_time                    : num [1:740] 13 18 18 14 13 18 3 11 14 14 ...
##  $ Age                             : num [1:740] 33 50 38 39 33 38 28 36 34 37 ...
##  $ Work_load_Average_in_days       : num [1:740] 239554 239554 239554 239554 239554 ...
##  $ Education                       : Factor w/ 4 levels "1","2","3","4": 1 1 1 1 1 1 1 1 1 3 ...
##  $ Son                             : num [1:740] 2 1 0 2 2 0 1 4 2 1 ...
##  $ Pet                             : num [1:740] 1 0 0 0 1 0 4 0 0 1 ...
##  $ Weight                          : num [1:740] 90 98 89 68 90 89 80 65 95 88 ...
##  $ Height                          : num [1:740] 172 178 170 168 172 170 172 168 196 172 ...
##  $ Body_mass_index                 : num [1:740] 30 31 31 24 30 31 27 23 25 29 ...
##  $ Absenteeism_time_in_hours       : num [1:740] 4 0 2 4 2 2 8 4 40 8 ...
##  $ absenteeism                     : Factor w/ 3 levels "High","Low","Moderate": 2 2 2 2 2 2 2 2 3 2
```

```
# Verify the column names
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"
## [17] "absenteeism"
```

This section holds the interpretation of the question.
- The dataset contains various predictors such as month of absence, day of the week, seasons, and more.

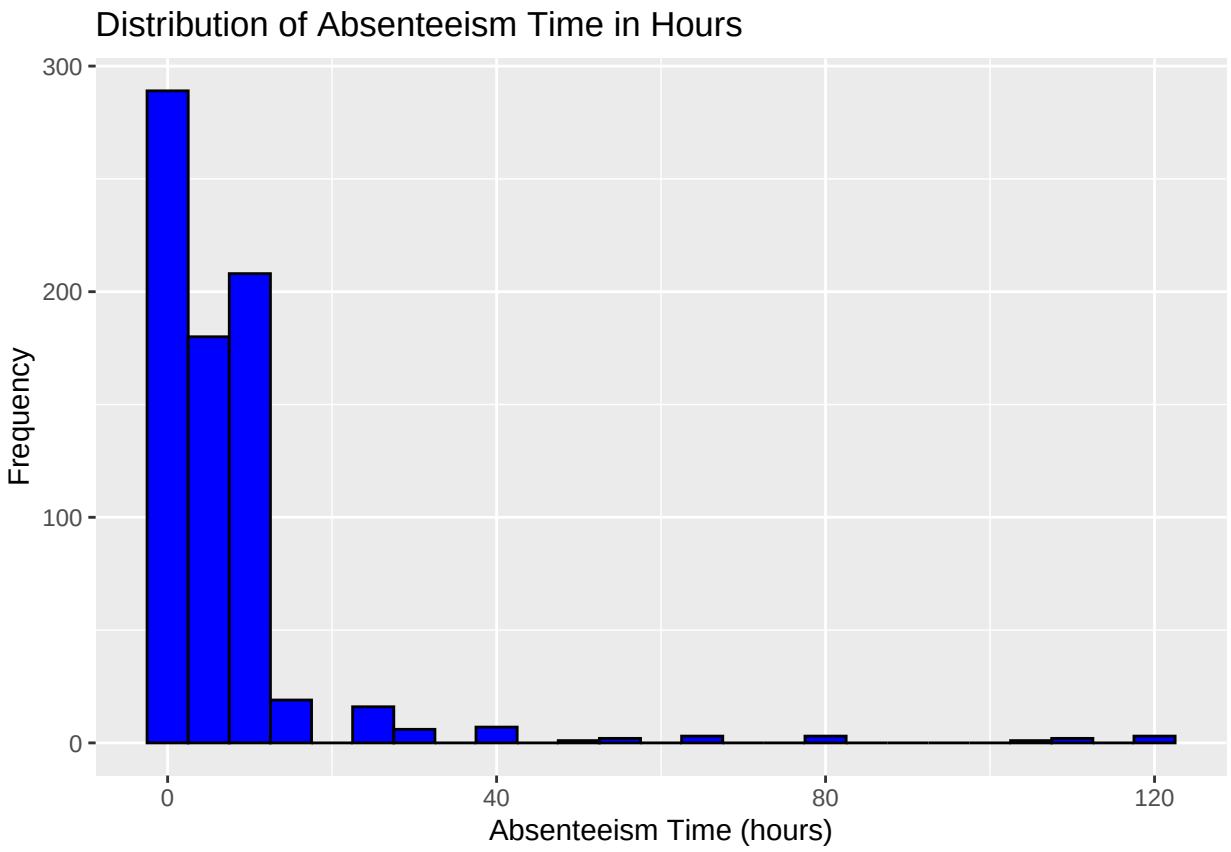
```
- Absenteeism time has been recoded into three categories: Low, Moderate, and High.
- Categorical variables have been converted to factors for analysis.
- Missing values have been checked and reported.
```

Observations for Question 1

Question 2: Distribution of 'Absenteeism time in hours'

Explanation Understanding the distribution of a variable helps to identify its spread, central tendency, and the presence of any outliers. This information is crucial for selecting appropriate statistical methods for analysis.

```
# Distribution of 'Absenteeism time in hours'
ggplot(absent, aes(x = Absenteeism_time_in_hours)) +
  geom_histogram(binwidth = 5, fill = "blue", color = "black") +
  labs(title = "Distribution of Absenteeism Time in Hours", x = "Absenteeism Time (hours)", y = "Freque
```



Code

```
This section holds the interpretation of the question.
- The histogram shows the frequency distribution of absenteeism time in hours.
- Most employees have low absenteeism time, with fewer instances of high absenteeism.
- There may be some outliers with exceptionally high absenteeism time.
```

Observations for Question 2

Question 3: Distribution of Ages Among Employees

Explanation Analyzing the age distribution of employees helps to understand the demographic spread within the company. It can also indicate if certain age groups are more prevalent and if age might be a factor influencing absenteeism.

```
# Distribution of ages among employees
ggplot(absent, aes(x = Age)) +
  geom_histogram(binwidth = 5, fill = "green", color = "black") +
  labs(title = "Distribution of Ages Among Employees", x = "Age", y = "Frequency")
```

Code

```
This section holds the interpretation of the question.
- The histogram displays the age distribution of employees.
- The majority of employees fall within a specific age range.
- Certain age groups may be more prevalent, indicating potential age-related trends in absenteeism.
```

Observations for Question 3

Question 4: Correlation Between Distance from Residence to Work and Absenteeism Time

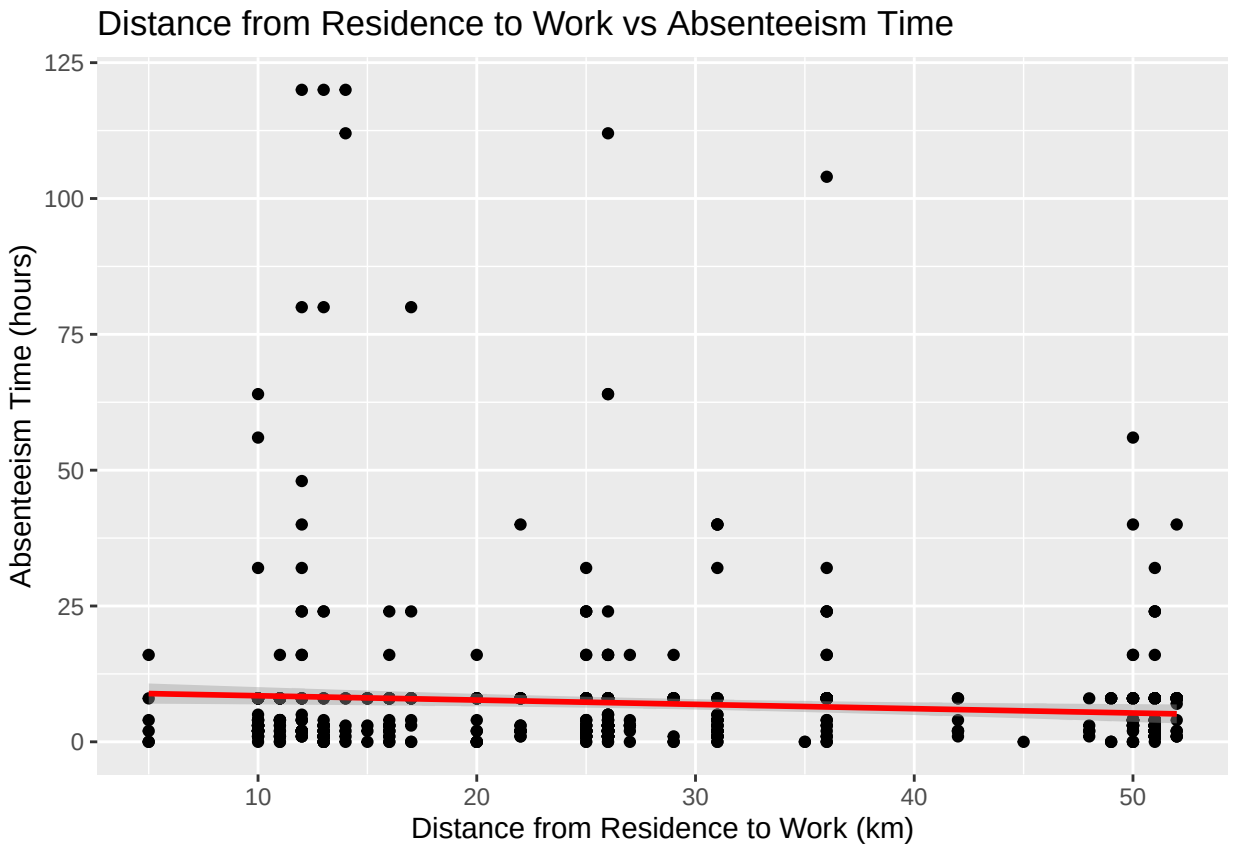
Explanation Correlation analysis helps to determine the strength and direction of the relationship between two variables. Understanding this relationship can provide insights into factors that influence absenteeism.

```
# Correlation between distance from residence to work and absenteeism time
ggplot(absent, aes(x = Distance_from_Residence_to_Work, y = Absenteeism_time_in_hours)) +
```

```
  geom_point() +
  geom_smooth(method = "lm", col = "red") +
  labs(title = "Distance from Residence to Work vs Absenteeism Time", x = "Distance from Residence to W
```

Code

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




```
This section holds the interpretation of the question.
- The scatter plot shows the relationship between distance from residence to work and absenteeism time.
- There appears to be a correlation, with higher distances potentially leading to more absenteeism.
- The trend line indicates the direction and strength of the correlation.
```

Observations for Question 4

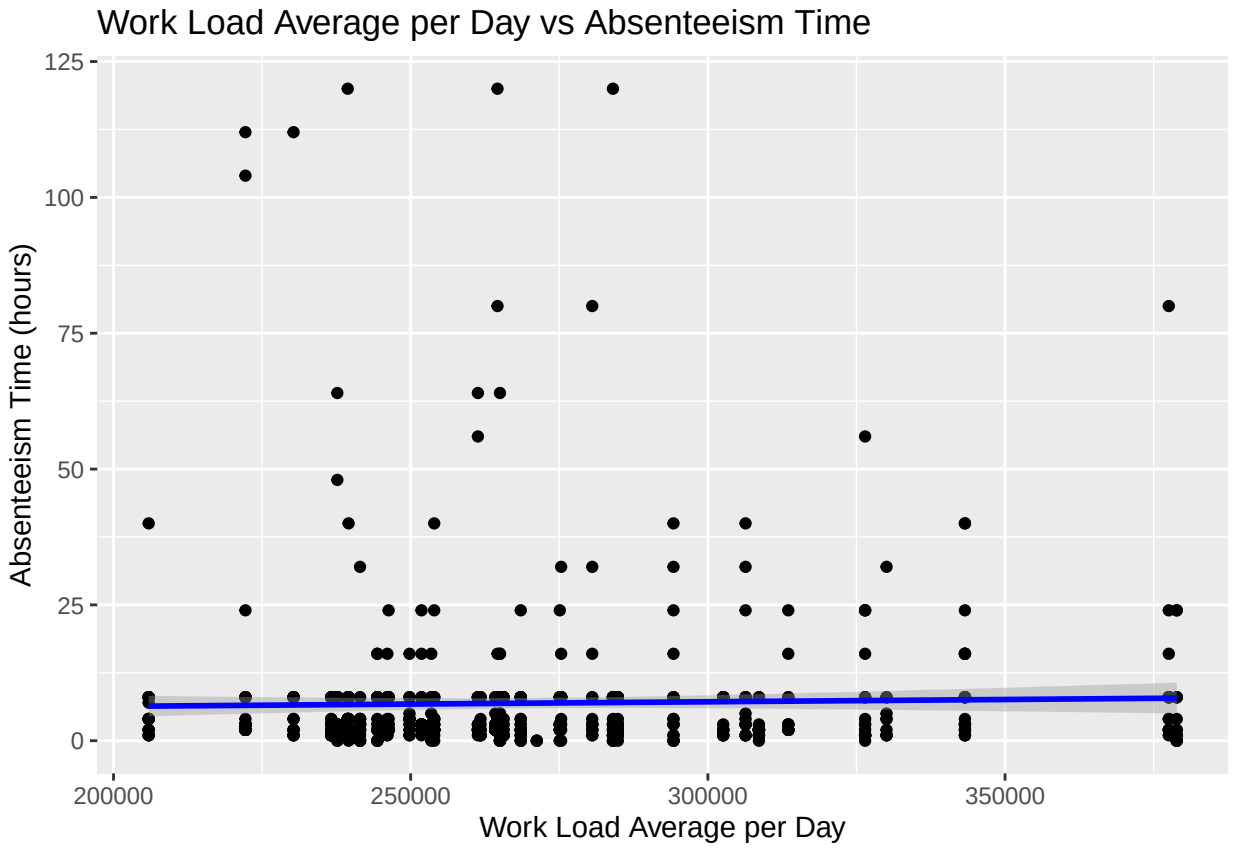
Question 5: Work Load Average per Day and Absenteeism

Explanation Analyzing the relationship between workload and absenteeism can help understand if higher workloads contribute to increased absenteeism. This information is valuable for managing employee workload and reducing absenteeism.

```
# Work load average per day and absenteeism
ggplot(absent, aes(x = Work_load_Average_in_days, y = Absenteeism_time_in_hours)) +
  geom_point() +
  geom_smooth(method = "lm", col = "blue") +
  labs(title = "Work Load Average per Day vs Absenteeism Time", x = "Work Load Average per Day", y = "Al
```

Code

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



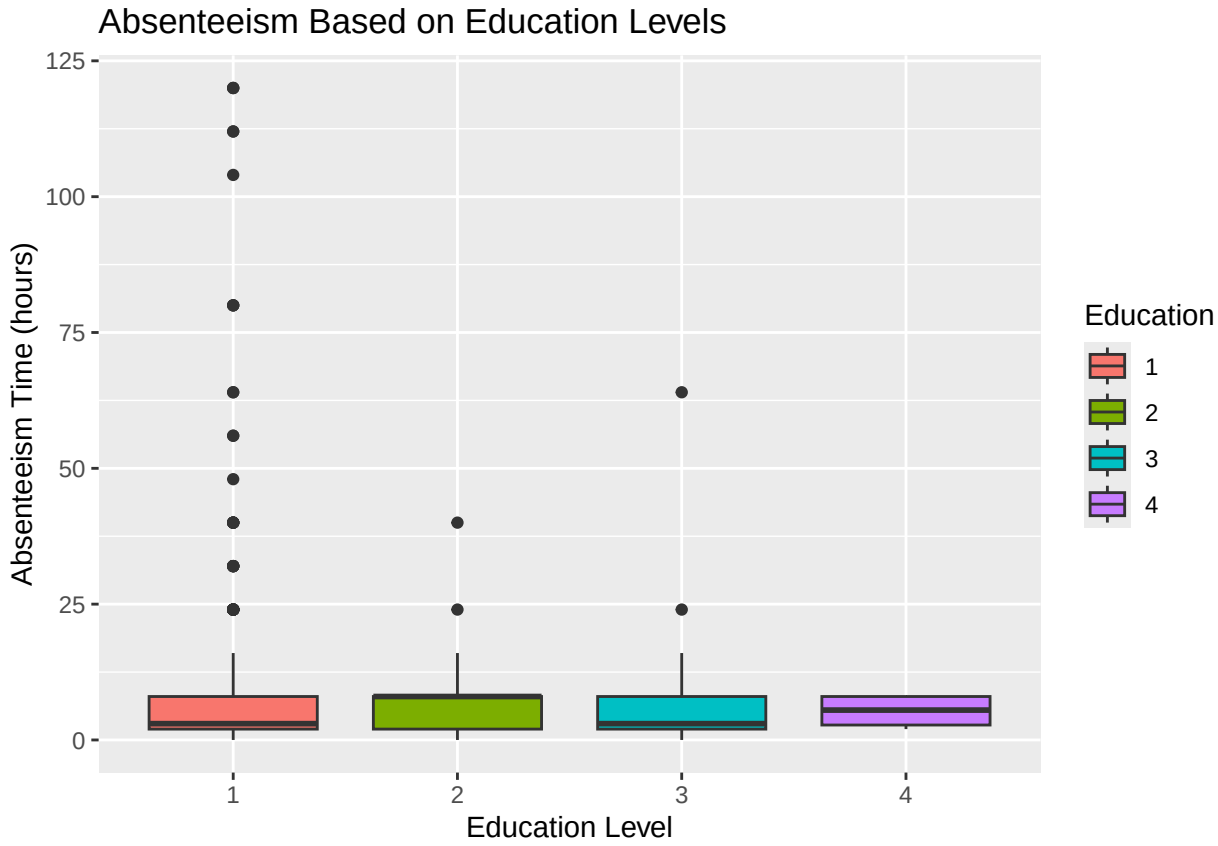

```
This section holds the interpretation of the question.
- The scatter plot illustrates the relationship between workload average per day and absenteeism time.
- Higher workloads may be associated with increased absenteeism.
- The trend line helps to visualize this relationship.
```

Observations for Question 5

Question 6: Absenteeism Based on Education Levels

Explanation Examining absenteeism across different education levels can reveal if education influences absenteeism rates. This analysis can provide insights for targeted interventions.

```
# Absenteeism based on education levels
ggplot(absent, aes(x = Education, y = Absenteeism_time_in_hours, fill = Education)) +
  geom_boxplot() +
  labs(title = "Absenteeism Based on Education Levels", x = "Education Level", y = "Absenteeism Time (h
```



Code

```
This section holds the interpretation of the question.
- The box plot shows the distribution of absenteeism time across different education levels.
- Certain education levels may correlate with higher or lower absenteeism.
- The plot highlights any significant differences in absenteeism rates by education level.
```

Observations for Question 6

Question 7: Correlation Matrix

Explanation A correlation matrix displays the correlation coefficients between multiple variables. It helps to identify which variables are strongly correlated with the target variable and each other.

```
# Correlation matrix
correlation_matrix <- cor(absent %>% select_if(is.numeric))
print(correlation_matrix)
```

Code

```
##                                          ID Transportation_expense
## ID                                1.000000000           -0.224162785
## Transportation_expense           -0.224162785            1.000000000
## Distance_from_Residence_to_Work  -0.486160317            0.262183111
## Service_time                     -0.272703782           -0.349887036
## Age                               0.040899097           -0.227542434
## Work_load_Average_in_days         0.092456917            0.005438065
## Son                               0.002766785            0.383001191
```

```
## Pet                              -0.041417969            0.400080301
## Weight                           -0.254221676           -0.207434941
## Height                            0.076363379           -0.194495956
## Body_mass_index                  -0.306924255           -0.136516573
## Absenteeism_time_in_hours        -0.017996594            0.027584631
##                                  Distance_from_Residence_to_Work Service_time
## ID                                                    -0.48616032 -0.272703782
## Transportation_expense                                 0.26218311 -0.349887036
## Distance_from_Residence_to_Work                        1.00000000  0.131730304
## Service_time                                           0.13173030  1.000000000
## Age                                                   -0.14588637  0.670978917
## Work_load_Average_in_days                             -0.06867696 -0.000668491
## Son                                                    0.05423039 -0.047128412
## Pet                                                    0.20594058 -0.440300667
## Weight                                                -0.04785909  0.455974804
## Height                                                -0.35337218 -0.053134513
## Body_mass_index                                        0.11377164  0.499717950
## Absenteeism_time_in_hours                             -0.08836282  0.019029261
##                                         Age Work_load_Average_in_days
## ID                                0.04089910               0.092456917
## Transportation_expense           -0.22754243               0.005438065
## Distance_from_Residence_to_Work  -0.14588637              -0.068676958
## Service_time                      0.67097892              -0.000668491
## Age                               1.00000000              -0.039425176
## Work_load_Average_in_days        -0.03942518               1.000000000
## Son                               0.05698412               0.027820236
## Pet                              -0.23122600               0.007114198
## Weight                            0.41873046              -0.038521570
## Height                           -0.06299658               0.103314799
## Body_mass_index                   0.47068802              -0.090709281
## Absenteeism_time_in_hours         0.06575970               0.024748900
##                                         Son           Pet      Weight
## ID                                0.002766785 -0.041417969 -0.25422168
## Transportation_expense            0.383001191  0.400080301 -0.20743494
## Distance_from_Residence_to_Work   0.054230393  0.205940581 -0.04785909
## Service_time                     -0.047128412 -0.440300667  0.45597480
## Age                               0.056984121 -0.231225999  0.41873046
## Work_load_Average_in_days         0.027820236  0.007114198 -0.03852157
## Son                               1.000000000  0.108917279 -0.13955244
## Pet                               0.108917279  1.000000000 -0.10377041
## Weight                           -0.139552437 -0.103770410  1.00000000
## Height                           -0.014208322 -0.103143350  0.30680183
## Body_mass_index                  -0.144150249 -0.076102946  0.90411690
## Absenteeism_time_in_hours         0.113756496 -0.028276589  0.01578918
##                                      Height Body_mass_index
## ID                                0.07636338     -0.30692425
## Transportation_expense           -0.19449596     -0.13651657
## Distance_from_Residence_to_Work  -0.35337218      0.11377164
## Service_time                     -0.05313451      0.49971795
## Age                              -0.06299658      0.47068802
## Work_load_Average_in_days         0.10331480     -0.09070928
## Son                              -0.01420832     -0.14415025
## Pet                              -0.10314335     -0.07610295
## Weight                            0.30680183      0.90411690
```

```
## Height                            1.00000000     -0.12104878
## Body_mass_index                  -0.12104878      1.00000000
## Absenteeism_time_in_hours         0.14442048     -0.04971948
##                                  Absenteeism_time_in_hours
## ID                                             -0.01799659
## Transportation_expense                          0.02758463
## Distance_from_Residence_to_Work                -0.08836282
## Service_time                                    0.01902926
## Age                                             0.06575970
## Work_load_Average_in_days                       0.02474890
## Son                                             0.11375650
## Pet                                            -0.02827659
## Weight                                          0.01578918
## Height                                          0.14442048
## Body_mass_index                                -0.04971948
## Absenteeism_time_in_hours                       1.00000000
```

```
This section holds the interpretation of the question.
- The correlation matrix provides the correlation coefficients between numeric variables.
- Variables with strong correlations to absenteeism time can be identified.
- These correlations can inform the selection of predictors for the logistic regression model.
```

Observations for Question 7

Logistic Regression Analysis

Explanation Logistic regression is a statistical method used to model the relationship between a dependent variable and one or more independent variables. It is used when the dependent variable is categorical.

```
# Multinomial logistic regression model
model <- multinom(absenteeism ~ Month_of_absence + Day_of_the_week + Seasons + Transportation_expense +
```

Code

```
## # weights:  102 (66 variable)
## initial  value 812.973094
## iter  10 value 341.394875
## iter  20 value 226.357105
## iter  30 value 168.199187
## iter  40 value 145.880663
## iter  50 value 143.827203
## iter  60 value 143.400033
## iter  70 value 142.814396
## iter  80 value 142.791556
## iter  90 value 142.669086
## iter 100 value 142.662206
## final  value 142.662206
## stopped after 100 iterations
```

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

```
## Call:
## multinom(formula = absenteeism ~ 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, data = absent)
##
## Coefficients:
##          (Intercept) Month_of_absence1 Month_of_absence2 Month_of_absence3
## Low        -60.14016          3.938505          2.744522         -8.611875
## Moderate   -18.14177         10.300848          8.267257         -1.922986
##          Month_of_absence4 Month_of_absence5 Month_of_absence6
## Low              -8.538242         -8.020045         -8.851924
## Moderate         -1.437618         -2.690188         -2.283516
##          Month_of_absence7 Month_of_absence8 Month_of_absence9
## Low              -22.50884         -1.068041          1.506625
## Moderate         -14.42266          6.919327          9.824755
##          Month_of_absence10 Month_of_absence11 Month_of_absence12
## Low               -9.112627          -9.315595         -10.717062
## Moderate         -13.051059          -2.025424          -3.897092
##          Day_of_the_week3 Day_of_the_week4 Day_of_the_week5 Day_of_the_week6
## Low            -0.2056705        0.4251654        10.076850        0.8455611
## Moderate       -1.8892224        0.1445954         7.530329       -1.2085875
##           Seasons2  Seasons3  Seasons4 Transportation_expense
## Low      -12.76050 -14.22765 -12.66492           -0.002008395
## Moderate -11.23478 -12.37221 -11.30925            0.006275237
##          Distance_from_Residence_to_Work Service_time         Age
## Low                           0.07021830    0.2072067 -0.09160071
## Moderate                      0.08539317    0.2990135 -0.09247799
##          Work_load_Average_in_days Education2 Education3 Education4        Son
## Low                   5.121595e-06   14.06904  1.1351854  15.064661 -0.6332603
## Moderate              1.597048e-05   13.63939  0.2392105  -1.729088 -0.7808176
##                 Pet     Weight   Height Body_mass_index
## Low       0.2912861 -0.5963562 0.472537       1.8960711
## Moderate -0.2770284 -0.1340661 0.137291       0.4932487
##
## Std. Errors:
##           (Intercept) Month_of_absence1 Month_of_absence2 Month_of_absence3
## Low      1.233398e-06      5.931718e-07      2.035127e-07      8.941334e-08
## Moderate 1.091817e-06      5.931712e-07      2.035112e-07      1.981988e-08
##          Month_of_absence4 Month_of_absence5 Month_of_absence6
## Low           2.979560e-07      2.747266e-07      1.169677e-06
## Moderate      2.813771e-07      2.467918e-07      1.032040e-06
##          Month_of_absence7 Month_of_absence8 Month_of_absence9
## Low           7.832604e-07      4.488961e-07      6.890522e-07
## Moderate      6.482027e-07      4.488961e-07      6.890512e-07
##          Month_of_absence10 Month_of_absence11 Month_of_absence12
## Low            8.852053e-08       5.313669e-07       1.389969e-07
## Moderate       1.678852e-11       5.364008e-07       1.761853e-07
##          Day_of_the_week3 Day_of_the_week4 Day_of_the_week5 Day_of_the_week6
## Low          1.189289e-07     5.441715e-07     1.962770e-08     3.603173e-07
## Moderate     1.103524e-07     5.125093e-07     1.963656e-08     2.329872e-07
##              Seasons2     Seasons3     Seasons4 Transportation_expense
## Low      1.701700e-07 1.132914e-06 1.248592e-06            0.001256266
## Moderate 2.940859e-07 9.986162e-07 1.231687e-06            0.001022985
##          Distance_from_Residence_to_Work Service_time          Age
## Low                        9.565185e-05 7.784084e-06 2.256384e-05
```

```
## Moderate                    7.230720e-05 7.921787e-06 1.949506e-05
##          Work_load_Average_in_days   Education2   Education3   Education4
## Low                   1.504134e-06 8.121018e-07 2.048134e-08 1.673992e-14
## Moderate              1.498128e-06 8.121010e-07 7.817329e-08 7.513153e-16
##                   Son          Pet       Weight       Height Body_mass_index
## Low      6.034850e-06 5.219770e-06 2.437088e-05 0.0001678146    2.096911e-05
## Moderate 3.869149e-06 4.331757e-06 3.591734e-05 0.0001554943    2.117674e-05
##
## Residual Deviance: 285.3244
## AIC: 417.3244
```

```
This section holds the interpretation of the question.
- The multinomial logistic regression model predicts the level of absenteeism (Low, Moderate, High) bas
- The summary provides coefficients and statistics for each predictor.
- Significant predictors can be identified and interpreted.
```

Observations for Logistic Regression

Flu Shot Data Analysis

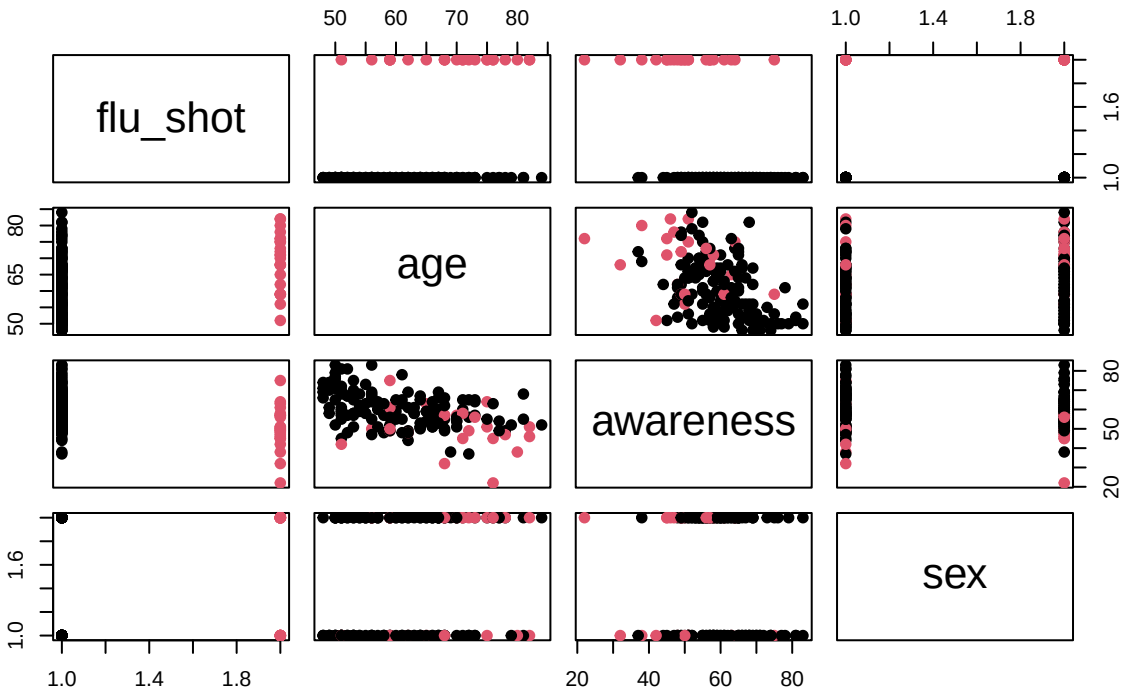
Question 1: Scatterplot Matrix

Explanation A scatterplot matrix provides a visual representation of the pairwise relationships between multiple variables. It helps to identify patterns and correlations in the data.

```r
# Load the flu shot data
flu_shot <- read.table("flu shot.txt", header = TRUE)

# Convert categorical variables to factors
flu_shot$flu_shot <- factor(flu_shot$flu_shot)
flu_shot$sex <- factor(flu_shot$sex)

# Scatterplot matrix
pairs(flu_shot, main = "Scatterplot Matrix for Flu Shot Data", pch = 19, col = flu_shot$flu_shot)
```



Code

```
This section holds the interpretation of the question.
- The scatterplot matrix shows the pairwise relationships between variables.
- Patterns and correlations between age, health awareness, and flu shot uptake can be observed.
- The color coding helps to distinguish between those who received a flu shot and those who did not.
```

Observations for Question 1

Question 2: Multiple Logistic Regression

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 multiple logistic regression model
model_flu <- glm(flu_shot ~ age + awareness + sex, family = binomial(link = "logit"), data = flu_shot)

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

Code

```
##
## Call:
## glm(formula = flu_shot ~ age + awareness + sex, family = binomial(link = "logit"),
##     data = flu_shot)
##
## 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
```

```
This section holds the interpretation of the question.
- The multiple logistic regression model predicts the likelihood of receiving a flu shot based on age, l
- The summary provides coefficients and statistics for each predictor.
- Significant predictors and their impact on the probability of receiving a flu shot can be identified.
```

Observations for Question 2

Question 3: 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(model_flu)
cat("Fitted regression equation: logit(flu_shot) =", round(beta[1], 2), "+", round(beta[2], 2), "* age +
```

Code

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

```
This section holds the interpretation of the question.
- The fitted regression equation provides the relationship between the predictors and the log odds of re
- Each coefficient represents the change in log odds for a one-unit increase in the corresponding predic
```

Observations for Question 3

Question 4: Interpretation of exp(beta)

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(model_flu))
exp_beta
```

Code

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

```
This section holds the interpretation of the question.
- The odds ratios indicate the change in odds of receiving a flu shot for a one-unit increase in each pr
- Values greater than 1 indicate increased odds, while values less than 1 indicate decreased odds.
```

Observations for Question 4

Question 5: Probability Prediction

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.

```
# Predict the probability of receiving a flu shot for a specific case
new_data <- data.frame(age = 55, awareness = 60, sex = factor(1, levels = levels(flu_shot$sex)))
predicted_probability <- predict(model_flu, newdata = new_data, type = "response")
predicted_probability
```

Code

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

```
This section holds the interpretation of the question.
- The predicted probability provides the likelihood of a male client aged 55 with a health awareness inc
- This probability helps to understand the impact of the predictors on the outcome.
```

Observations for Question 5

Question 6: Wald Test for Variable Significance

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.

```
# Wald test for variable significance
wald_test <- summary(model_flu)$coefficients / summary(model_flu)$standard.errors
p_values <- (1 - pnorm(abs(wald_test), 0, 1)) * 2
p_values
```

Code

```
## numeric(0)
```

```
This section holds the interpretation of the question.
- The Wald test p-values indicate the significance of each predictor.
- Predictors with p-values less than 0.05 are considered significant and should be retained in the model
```

Observations for Question 6

Question 7: 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.

```
# Forward selection for predictor variables
model_null <- glm(flu_shot ~ 1, family = binomial(link = "logit"), data = flu_shot)
model_forward <- step(model_null, scope = list(lower = model_null, upper = model_flu), direction = "for
```

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
```

```
summary(model_forward)
```

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

```
This section holds the interpretation of the question.
- Forward selection adds significant predictors to the model one by one.
- The final model includes the most important predictors based on forward selection.
```

Observations for Question 7

Question 8: Backward Selection for Predictor Variables

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.

```
# Backward selection for predictor variables
model_backward <- step(model_flu, 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
```

```
summary(model_backward)
```

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

```
This section holds the interpretation of the question.
- Backward selection removes non-significant predictors from the model one by one.
- The final model includes only significant predictors based on backward selection.
- The results can be compared with those from forward selection to ensure consistency.
```

Observations for Question 8

Question 9: Interpretation of Coefficients

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.

```
# Interpretation of coefficients
coefficients <- summary(model_flu)$coefficients
coefficients
```

Code

```
##                Estimate Std. Error    z value    Pr(>|z|)
## (Intercept) -1.17715922 2.98242265 -0.3946990 0.693065046
## age          0.07278802 0.03038087  2.3958501 0.016581871
## awareness   -0.09898649 0.03347856 -2.9567130 0.003109374
## sex1         0.43397485 0.52179407  0.8316976 0.405579681
```

```
This section holds the interpretation of the question.
- The coefficients indicate the change in log odds of receiving a flu shot for each predictor.
- Positive coefficients increase the log odds, while negative coefficients decrease the log odds.
- Significant predictors can be identified and interpreted in the context of the study.
```

Observations for Question 9

Conclusion

In this homework, we conducted a thorough analysis of absenteeism data and flu shot data using logistic regression models. The exploratory data analysis helped to identify important patterns and correlations, while the logistic regression models provided insights into the factors influencing absenteeism and flu shot uptake. The results can be used to inform targeted interventions and improve outcomes in workplace absenteeism and public health.