

Final

Orator

2025-11-21

Data Selection and Dataset Description

For this project, I selected the OSMI Mental Health in Tech Survey dataset, which contains self-reported mental-health information from individuals working in the technology industry. The dataset was collected by Open Sourcing Mental Illness (OSMI) to better understand how workplace conditions, demographics, and employer policies influence mental-health outcomes among tech workers.

The dataset includes 1,259 observations and over 25 variables covering demographic characteristics (Age, Gender, Country, state), mental-health history (family_history, treatment), and workplace factors such as company size (no_employees), remote work, employer-provided benefits, wellness programs, and leave policies. Several variables also capture perceptions about workplace stigma, such as comfort discussing mental health with supervisors, coworkers, or during job interviews, and concerns about negative consequences.

In this project, my goal is to examine how personal characteristics and workplace environments relate to mental-health outcomes. I will construct a continuous Mental Health Risk Score using several mental-health-related variables. This constructed score will serve as the response variable for the linear regression model. Predictor variables will include demographic features, workplace characteristics, and cultural factors related to mental-health openness.

Loading Dataset

```
survey <- read.csv("survey.csv", stringsAsFactors = FALSE)
str(survey)

## 'data.frame': 1259 obs. of 27 variables:
## $ Timestamp: chr "2014-08-27 11:29:31" "2014-08-27 11:29:37" "2014-08-27 11:29:44"
## $ Age      : num 37 44 32 31 31 33 35 39 42 23 ...
## $ Gender   : chr "Female" "M" "Male" "Male" ...
## $ Country  : chr "United States" "United States" "Canada" "United Kingdom" ...
## $ state    : chr "IL" "IN" NA NA ...
## $ self_employed: chr NA NA NA NA ...
## $ family_history: chr "No" "No" "No" "Yes" ...
## $ treatment: chr "Yes" "No" "No" "Yes" ...
## $ work_interfere: chr "Often" "Rarely" "Rarely" "Often" ...
## $ no_employees: chr "6-25" "More than 1000" "6-25" "26-100" ...
## $ remote_work: chr "No" "No" "No" "No" ...
## $ tech_company: chr "Yes" "No" "Yes" "Yes" ...
## $ benefits: chr "Yes" "Don't know" "No" "No" ...
## $ care_options: chr "Not sure" "No" "No" "Yes" ...
## $ wellness_program: chr "No" "Don't know" "No" "No" ...
## $ seek_help: chr "Yes" "Don't know" "No" "No" ...
## $ anonymity: chr "Yes" "Don't know" "Don't know" "No" ...
## $ leave: chr "Somewhat easy" "Don't know" "Somewhat difficult" "Somewhat diffi
```

```

## $ mental_health_consequence: chr  "No" "Maybe" "No" "Yes" ...
## $ phys_health_consequence : chr  "No" "No" "No" "Yes" ...
## $ coworkers              : chr  "Some of them" "No" "Yes" "Some of them" ...
## $ supervisor              : chr  "Yes" "No" "Yes" "No" ...
## $ mental_health_interview: chr  "No" "No" "Yes" "Maybe" ...
## $ phys_health_interview   : chr  "Maybe" "No" "Yes" "Maybe" ...
## $ mental_vs_physical      : chr  "Yes" "Don't know" "No" "No" ...
## $ obs_consequence         : chr  "No" "No" "No" "Yes" ...
## $ comments                : chr  NA NA NA NA ...

head(survey)

##           Timestamp Age Gender          Country state self_employed
## 1 2014-08-27 11:29:31  37 Female United States    IL        <NA>
## 2 2014-08-27 11:29:37  44      M United States    IN        <NA>
## 3 2014-08-27 11:29:44  32    Male Canada       <NA>        <NA>
## 4 2014-08-27 11:29:46  31    Male United Kingdom <NA>        <NA>
## 5 2014-08-27 11:30:22  31    Male United States   TX        <NA>
## 6 2014-08-27 11:31:22  33    Male United States   TN        <NA>
## family_history treatment work_interfere no_employees remote_work
## 1          No      Yes        Often      6-25        No
## 2          No      No        Rarely More than 1000     No
## 3          No      No        Rarely      6-25        No
## 4          Yes     Yes        Often      26-100      No
## 5          No      No        Never     100-500      Yes
## 6          Yes     No        Sometimes      6-25        No
## tech_company benefits care_options wellness_program seek_help anonymity
## 1          Yes     Yes      Not sure        No      Yes      Yes
## 2          No Don't know        No      Don't know Don't know Don't know
## 3          Yes     No        No        No      No Don't know
## 4          Yes     No        Yes        No      No      No
## 5          Yes     Yes        No      Don't know Don't know Don't know
## 6          Yes     Yes      Not sure        No      Don't know Don't know
## leave mental_health_consequence phys_health_consequence
## 1 Somewhat easy        No        No
## 2 Don't know           Maybe      No
## 3 Somewhat difficult      No        No
## 4 Somewhat difficult      Yes      Yes
## 5 Don't know           No        No
## 6 Don't know           No        No
## coworkers supervisor mental_health_interview phys_health_interview
## 1 Some of them      Yes        No      Maybe
## 2          No        No        No        No
## 3          Yes     Yes        Yes      Yes
## 4 Some of them      No        Maybe      Maybe
## 5 Some of them      Yes        Yes      Yes
## 6          Yes     Yes        No      Maybe
## mental_vs_physical obs_consequence comments
## 1          Yes        No      <NA>
## 2 Don't know           No      <NA>
## 3          No        No      <NA>
## 4          No        Yes      <NA>
## 5 Don't know           No      <NA>
## 6 Don't know           No      <NA>

```

Descriptive Statistics

```
survey <- subset(survey, Age >= 15 & Age <= 80)

summary(survey$Age)

##      Min. 1st Qu. Median     Mean 3rd Qu.     Max.
## 18.00    27.00   31.00   32.08   36.00   72.00

table(survey$treatment)

##
##  No Yes
## 619 632

table(survey$family_history)

##
##  No Yes
## 762 489

table(survey$work_interfere, useNA = "ifany")

##
##      Never      Often      Rarely Sometimes      <NA>
##        212       140       173       464       262

table(survey$benefits)

##
##  Don't know      No      Yes
##        407       371       473

summary(survey)

##      Timestamp            Age          Gender          Country
##  Length:1251   Min.   :18.00  Length:1251   Length:1251
##  Class :character 1st Qu.:27.00  Class :character  Class :character
##  Mode  :character Median :31.00   Mode  :character  Mode  :character
##                                         Mean   :32.08
##                                         3rd Qu.:36.00
##                                         Max.   :72.00
##      state           self-employed family_history treatment
##  Length:1251   Length:1251    Length:1251    Length:1251
##  Class :character Class :character Class :character  Class :character
##  Mode  :character Mode  :character Mode  :character  Mode  :character
## 
## 
##      work_interfere      no_employees      remote_work      tech_company
##  Length:1251   Length:1251    Length:1251    Length:1251
##  Class :character Class :character Class :character  Class :character
##  Mode  :character Mode  :character Mode  :character  Mode  :character
## 
## 
##      benefits         care_options      wellness_program      seek_help
##  Length:1251   Length:1251    Length:1251    Length:1251
##  Class :character Class :character Class :character  Class :character
##  Mode  :character Mode  :character Mode  :character  Mode  :character
```

```

##  Length:1251      Length:1251      Length:1251      Length:1251
##  Class :character Class :character Class :character Class :character
##  Mode  :character Mode  :character Mode  :character Mode  :character
##
## 
## 
## 
##  anonymity          leave        mental_health_consequence
##  Length:1251      Length:1251      Length:1251
##  Class :character Class :character Class :character
##  Mode  :character Mode  :character Mode  :character
##
## 
## 
## 
##  phys_health_consequence coworkers supervisor
##  Length:1251      Length:1251      Length:1251
##  Class :character Class :character Class :character
##  Mode  :character Mode  :character Mode  :character
##
## 
## 
## 
##  mental_health_interview phys_health_interview mental_vs_physical
##  Length:1251      Length:1251      Length:1251
##  Class :character Class :character Class :character
##  Mode  :character Mode  :character Mode  :character
##
## 
## 
## 
##  obs_consequence comments
##  Length:1251      Length:1251
##  Class :character Class :character
##  Mode  :character Mode  :character
##
## 
## 
## 
```

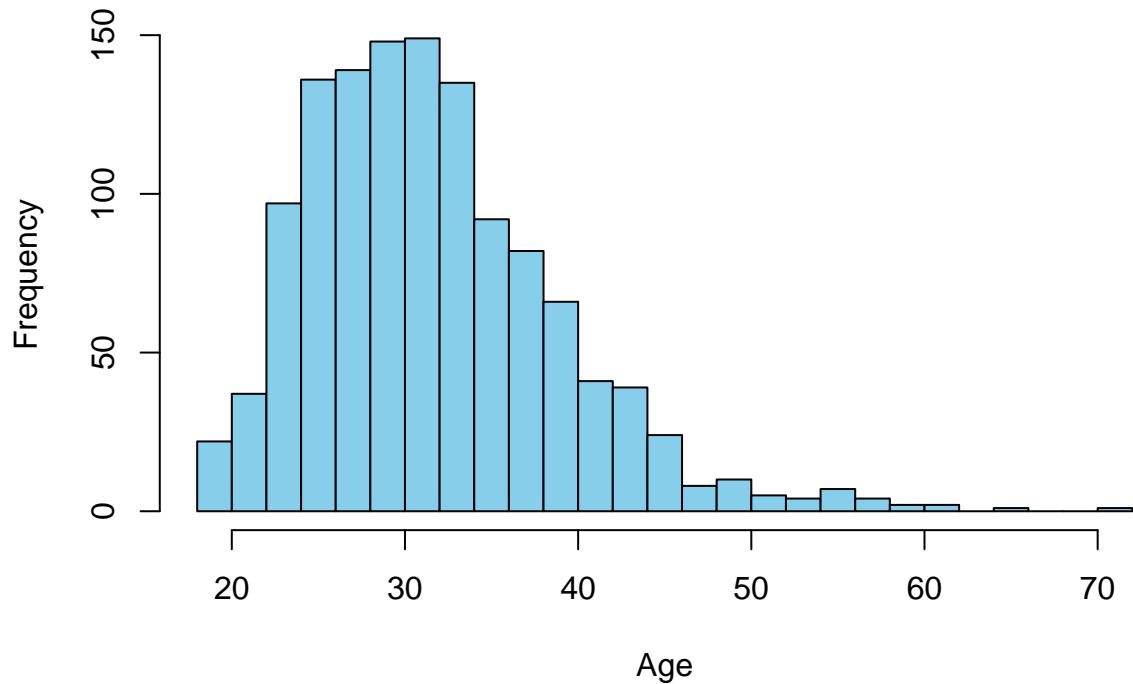
Visualizations for Data Exploration

```

# Histogram of Age
hist(survey$Age,
      main = "Distribution of Age",
      xlab = "Age",
      col = "skyblue",
      breaks = 20)

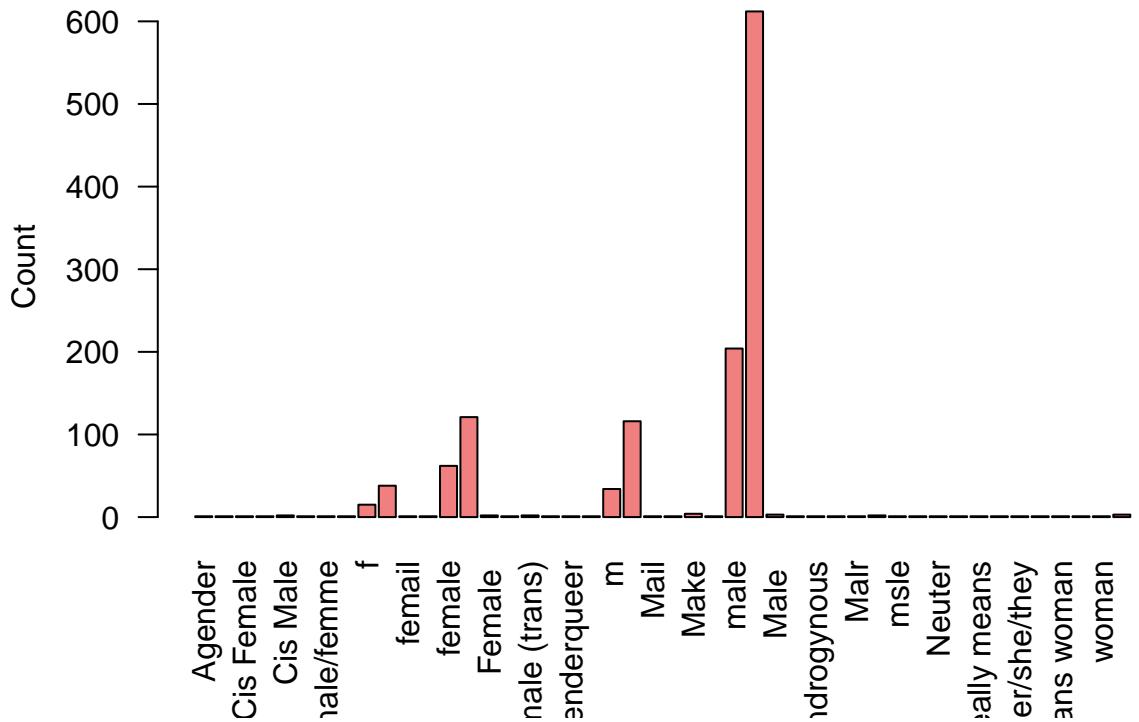
```

Distribution of Age



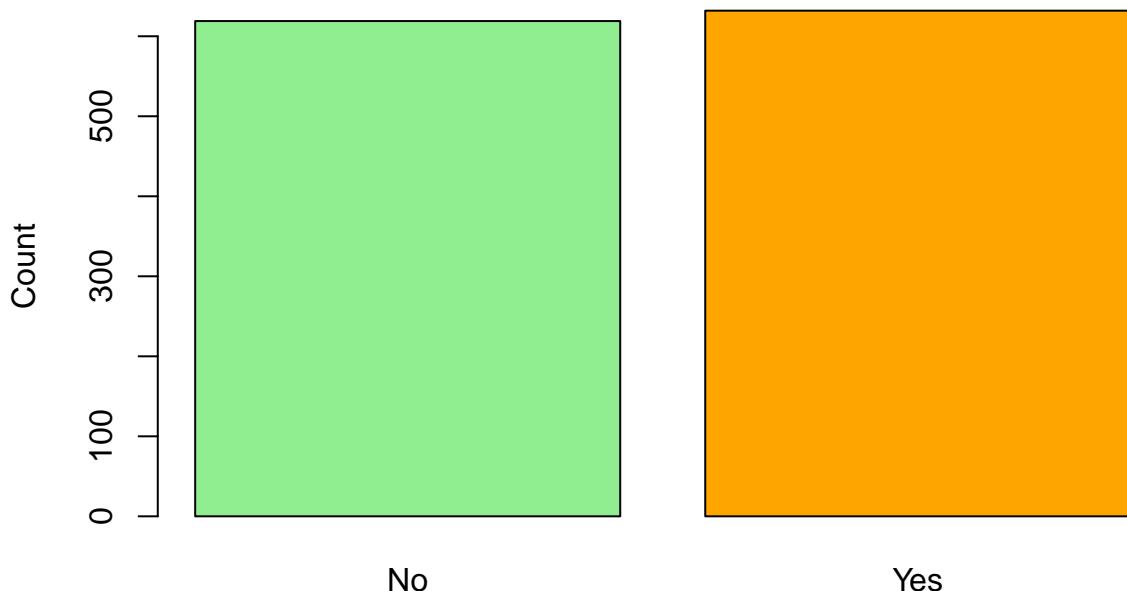
```
# Bar Plot for Gender Distribution
gender_tab <- table(survey$Gender)
barplot(gender_tab,
        main = "Gender Distribution",
        ylab = "Count",
        col = "lightcoral",
        las = 2)
```

Gender Distribution



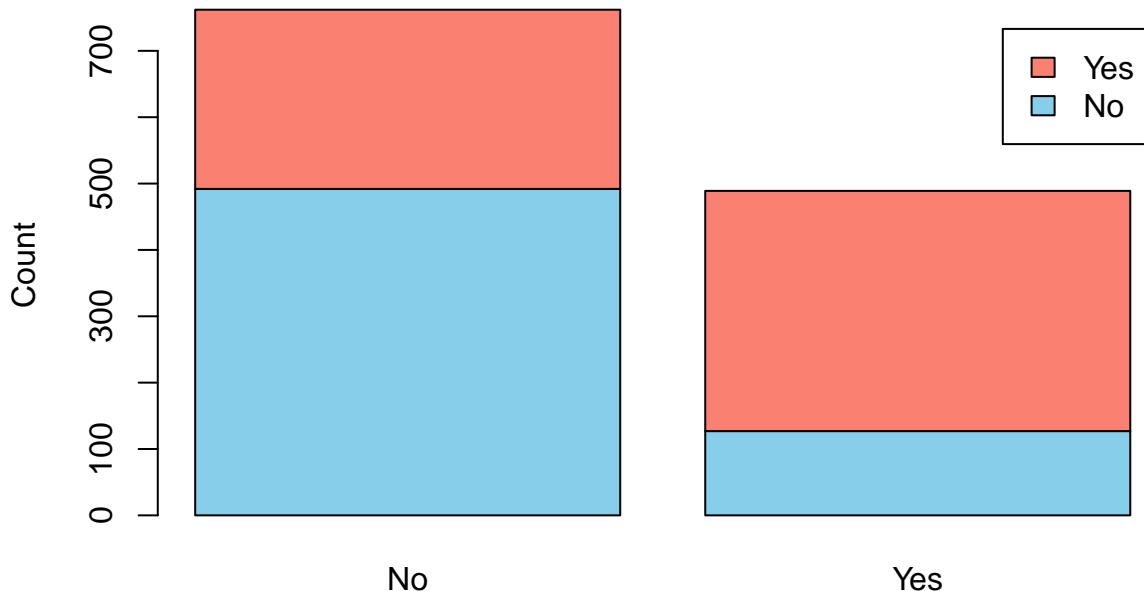
```
# Bar plot of Treatment (Yes/No)
treat_tab <- table(survey$treatment)
barplot(treat_tab,
        main = "Mental Health Treatment",
        ylab = "Count",
        col = c("lightgreen", "orange"))
```

Mental Health Treatment



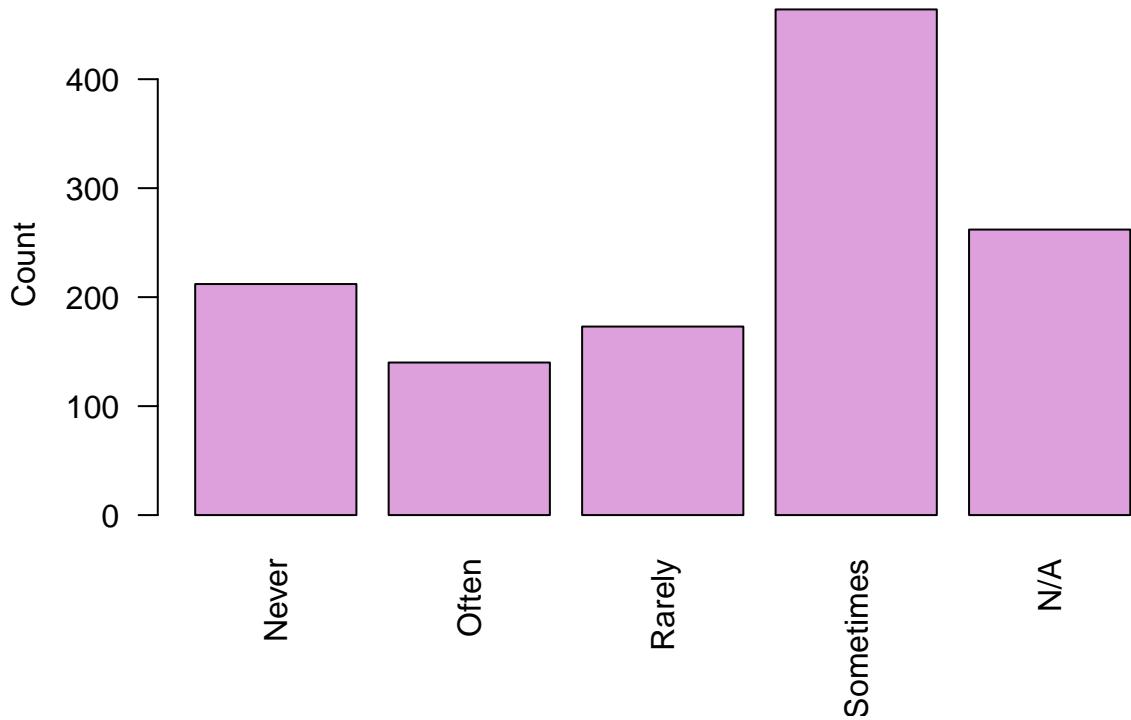
```
#Stacked bar Plot for treatment by family history
tab_treat_hist <- table(survey$treatment, survey$family_history)
barplot(tab_treat_hist,
        main = "Treatment by Family History",
        ylab = "Count",
        col = c("skyblue", "salmon"),
        legend = rownames(tab_treat_hist))
```

Treatment by Family History



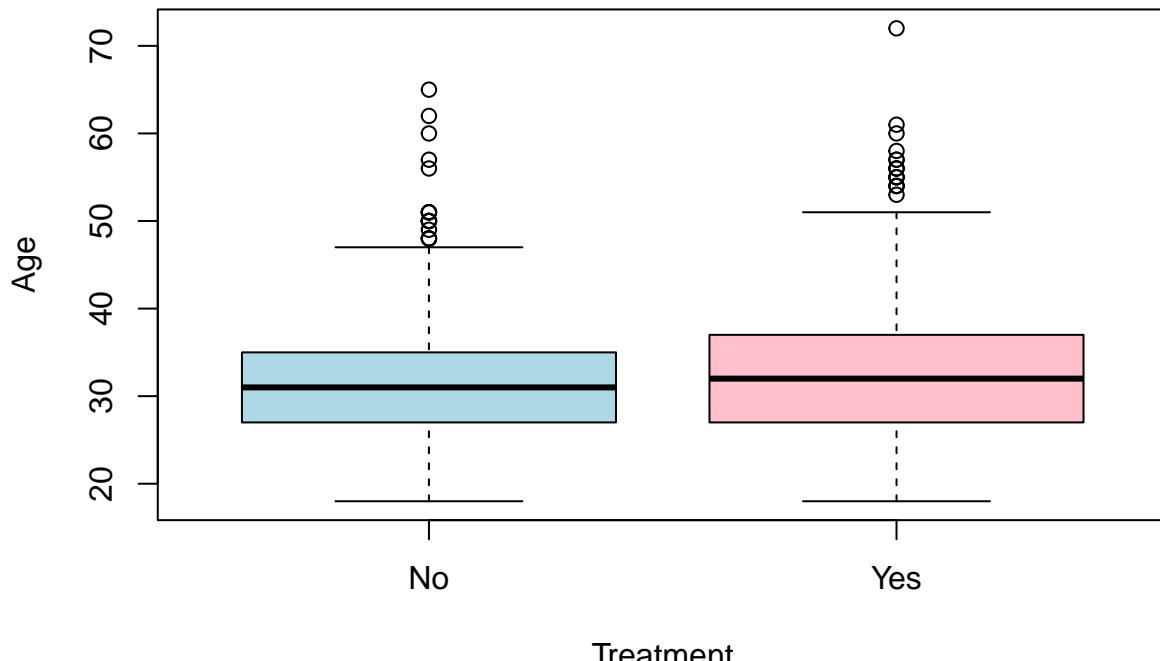
```
# Bar plot of Work Interference
work_tab <- table(survey$work_interfere, useNA = "ifany")
names(work_tab)[is.na(names(work_tab))] <- "N/A"
barplot(work_tab,
        main = "How Mental Health Interferes With Work",
        ylab = "Count",
        las = 2,
        col = "plum")
```

How Mental Health Interferes With Work



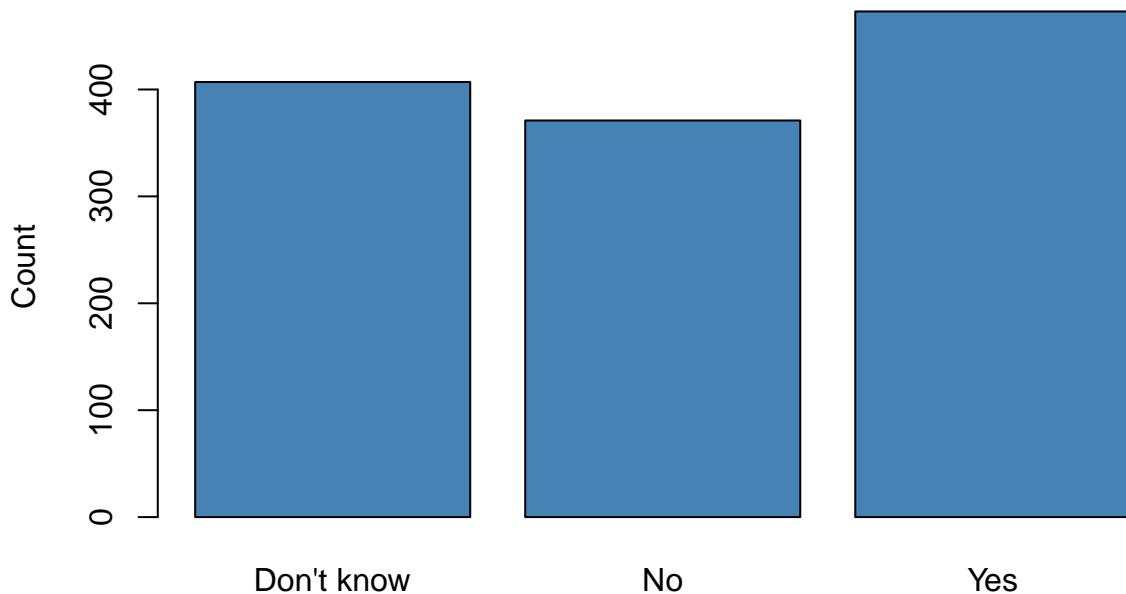
```
# Boxplot: Age vs Treatment
boxplot(Age ~ treatment, data = survey,
        main = "Age Distribution by Treatment Status",
        xlab = "Treatment",
        ylab = "Age",
        col = c("lightblue", "pink"))
```

Age Distribution by Treatment Status



```
#bar plot for availability of mental health benefits
benefits_tab <- table(survey$benefits)
barplot(benefits_tab,
        main = "Employer Mental Health Benefits",
        ylab = "Count",
        col = "steelblue")
```

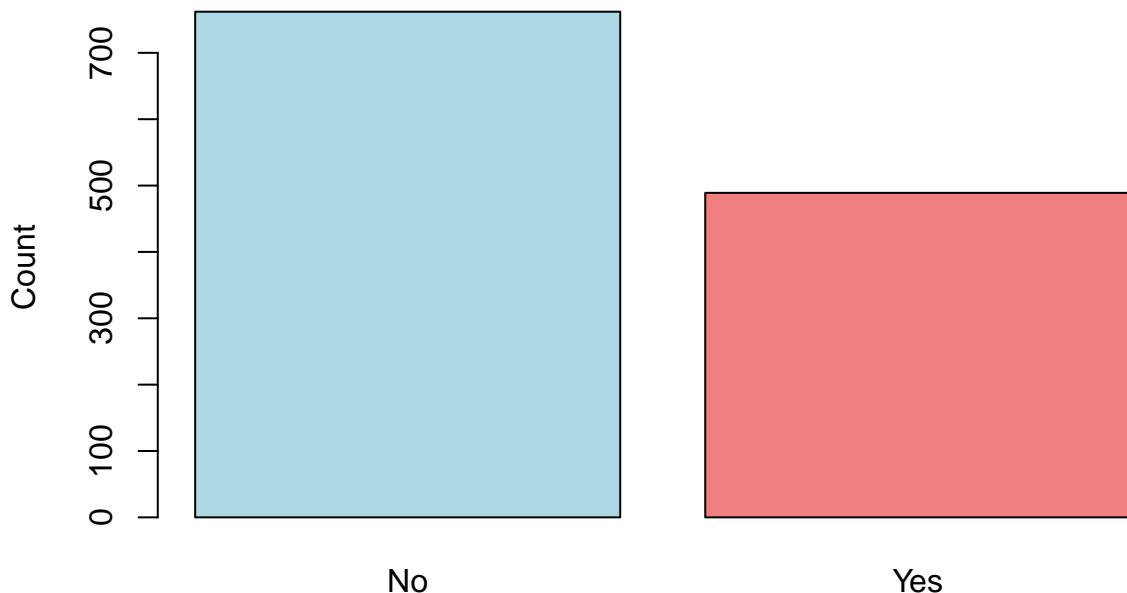
Employer Mental Health Benefits



```
# Bar Plot: Family history
fam_tab <- table(survey$family_history, useNA = "ifany")

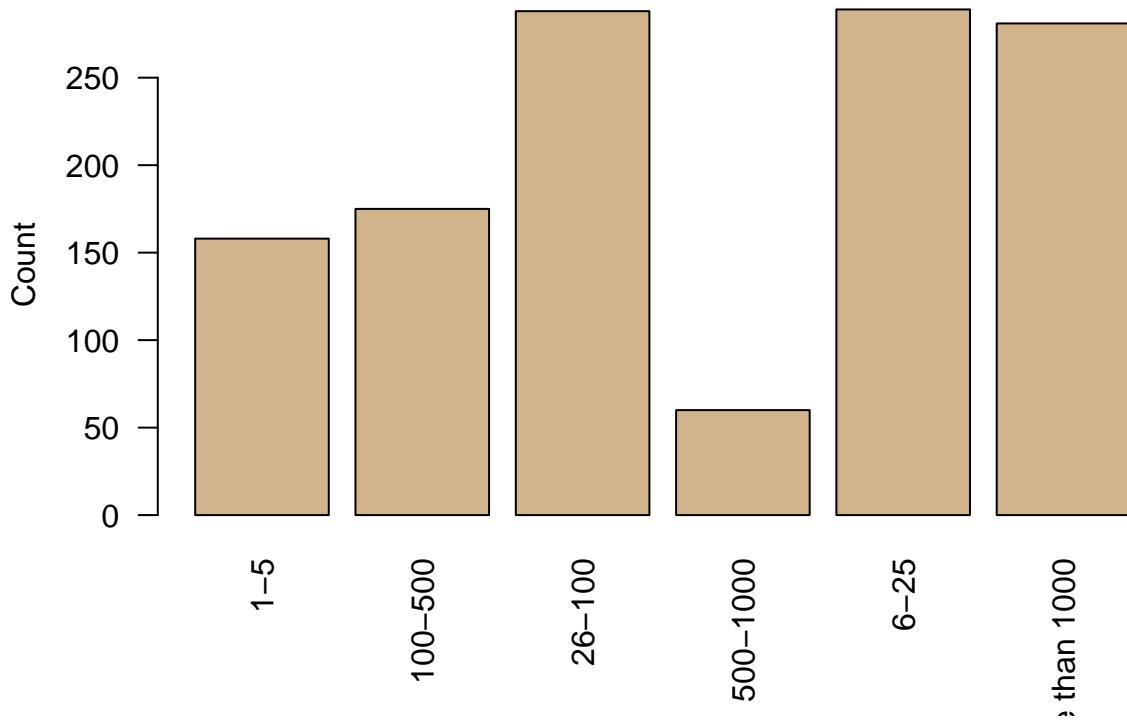
barplot(fam_tab,
        main = "Family History of Mental Illness",
        ylab = "Count",
        col = c("lightblue", "lightcoral"))
```

Family History of Mental Illness



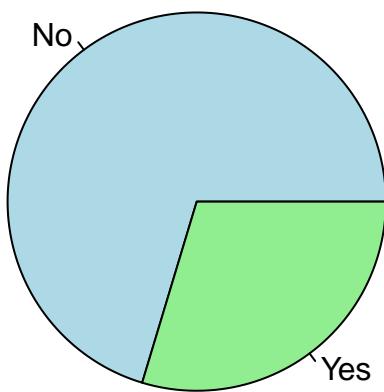
```
#Bar Plot: Company Size Distribution
size_tab <- table(survey$no_employees)
barplot(size_tab,
        main = "Company Size Distribution",
        ylab = "Count",
        las = 2,
        col = "tan")
```

Company Size Distribution



```
#Pie Chart: Remote Work
remote_tab <- table(survey$remote_work)
pie(remote_tab,
  main = "Remote Work Distribution",
  col = c("lightblue", "lightgreen"))
```

Remote Work Distribution

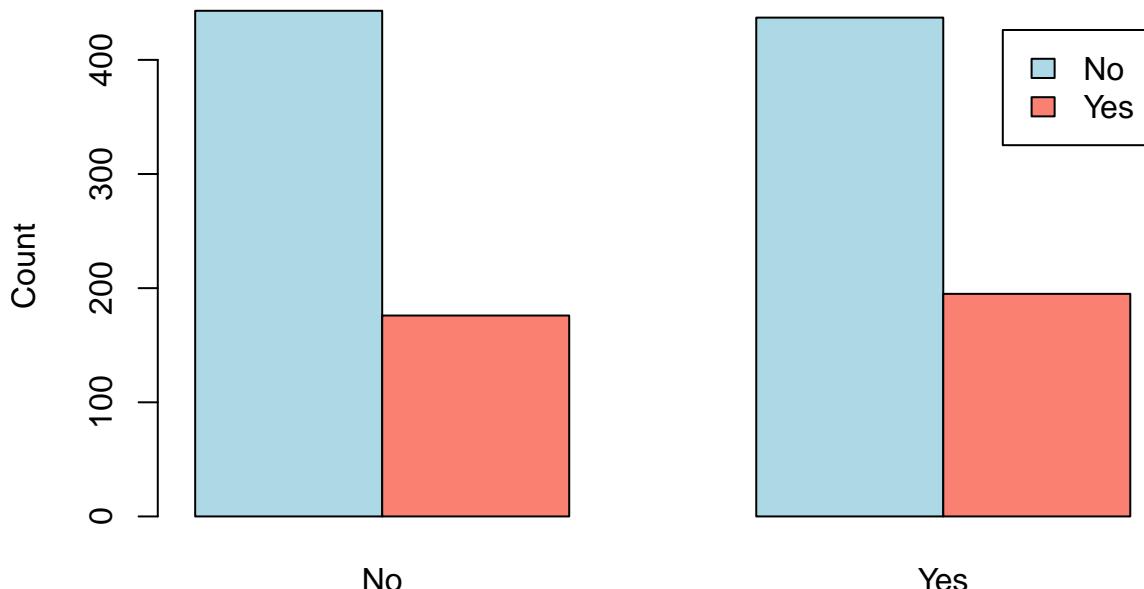


```
#Barplot: Remote Work vs Treatment
remote_treat <- table(survey$remote_work, survey$treatment)

rownames(remote_treat)[is.na(rownames(remote_treat))] <- "NA"
```

```
barplot(remote_treat,
        main = "Remote Work vs Treatment Status",
        ylab = "Count",
        col = c("lightblue", "salmon"),
        beside = TRUE,
        legend = colnames(remote_treat))
```

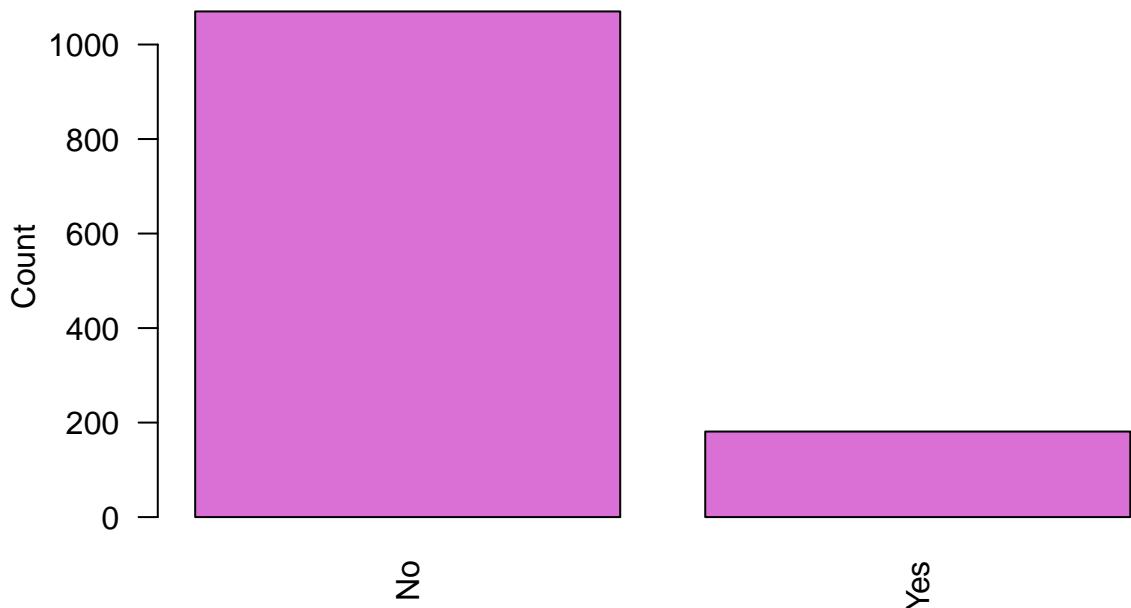
Remote Work vs Treatment Status



```
# Bar Plot: Observed consequences
obs_tab <- table(survey$obs_consequence, useNA = "ifany")

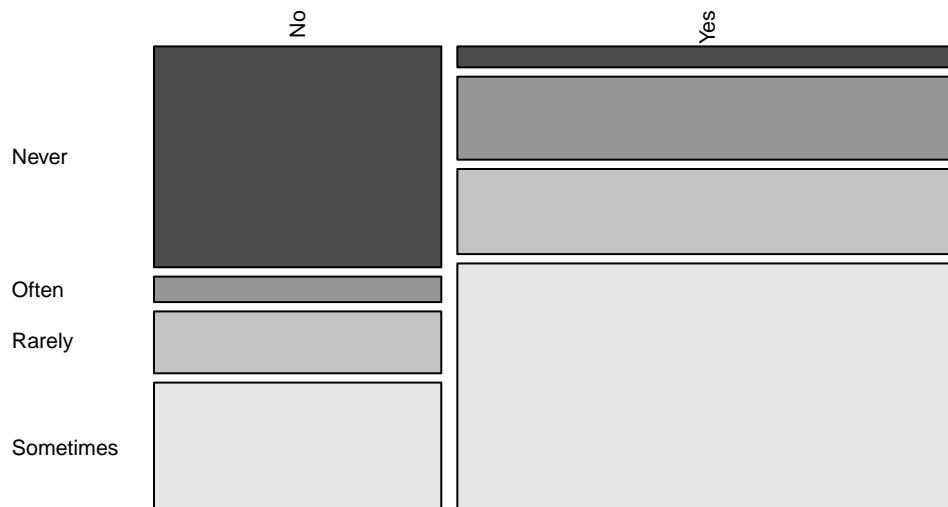
barplot(obs_tab,
        main = "Observed Workplace Consequences Related to Mental Health",
        ylab = "Count",
        las = 2,
        col = "orchid")
```

Observed Workplace Consequences Related to Mental Health



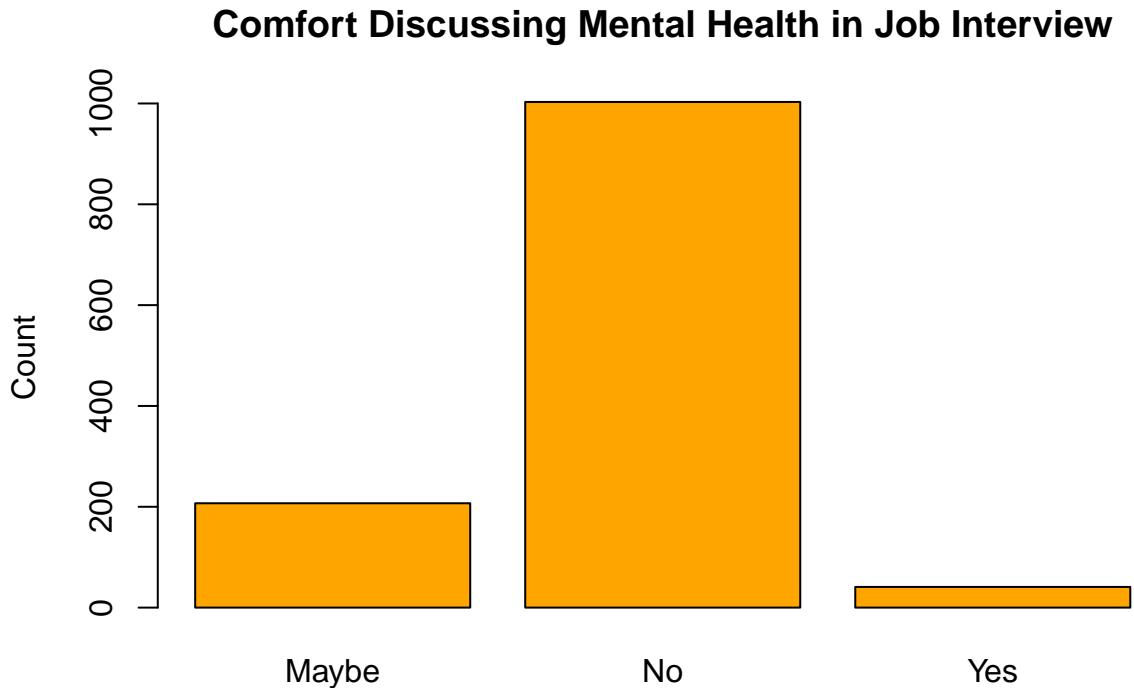
```
#Mosaic Plot: Treatment vs Work Interference
table_tw <- table(survey$treatment, survey$work_interfere)
mosaicplot(table_tw,
           main = "Treatment vs Work Interference",
           color = TRUE,
           las = 2)
```

Treatment vs Work Interference



```
#Bar Plot: Mental Health Interview Comfort
inter_tab <- table(survey$mental_health_interview)
barplot(inter_tab,
        main = "Comfort Discussing Mental Health in Job Interview",
        ylab = "Count",
```

```
col = "orange")
```



Data Visualisation Summary and Insights

The age distribution shows that most respondents are in their mid-20s to mid-30s, with fewer older participants. Gender entries are diverse, but “Male” and “Female” dominate. Mental health treatment is nearly evenly split, though people with a family history of mental illness are noticeably more likely to receive treatment. Work interference varies, but “Sometimes” is the most common response, showing that many employees are affected occasionally.

Company sizes and remote-work patterns are mixed, with most respondents not working remotely. Many employees are unsure whether their employer provides mental health benefits, though a significant portion does receive them. In terms of attitudes, most respondents feel uncomfortable discussing mental health during job interviews, highlighting strong stigma. Few people have observed negative workplace consequences, but those who have are far more likely to report interference and treatment.

Visuals comparing treatment status to work interference show that people who experience symptoms more often are more likely to seek help. Remote work does not substantially change treatment patterns. Generally, the charts suggest that while mental health concerns are common and sometimes affect job performance, stigma remains high, especially in hiring and access to or awareness of employer support is inconsistent across workplaces.

Response Variable and Predictor Variables

Response Variable

For this project, I define the Mental Health Risk Score as the response variable I want to predict.

Predictor Variables

Potential predictor variables that may influence mental-health risk include:

- Age
- Gender
- Country
- Remote work status
- Company size
- Family history of mental illness
- Employer benefits
- Availability of care options
- Supervisor support
- Workplace culture indicators such as:
 - anonymity protections
 - ease of taking mental-health leave
 - comfort discussing mental health (mental_health_interview)
 - perceived consequences

Data Preprocessing

```
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

library(stringr)

#Removing irrelevant columns
survey <- survey %>% select(!any_of(c("Timestamp", "comments")))

# Handling missing values

#Work Interference → Replacing NA with "Unknown"

survey$work_interfere[survey$work_interfere == "" | 
  is.na(survey$work_interfere)] <- "Unknown"

#removing other high na columns
survey <- survey %>% select(-state, -self_employed)

#cleaning Gender into Male, Female and Other
survey <- survey %>%
  mutate(
    Gender = str_to_lower(Gender),
    Gender = case_when(
      str_detect(Gender, "f") ~ "Female",
      str_detect(Gender, "m") ~ "Male",
```

```

    TRUE ~ "Other"
),
Gender = factor(Gender)
)
survey$Gender <- factor(survey$Gender)
table(survey$Gender)

## 
## Female   Male   Other
##    247     994     10

#Converting relevant character varibales to factors
categorical_vars <- c(
  "treatment", "family_history", "remote_work", "tech_company",
  "benefits", "care_options", "wellness_program", "seek_help",
  "anonymity", "leave", "mental_health_consequence",
  "phys_health_consequence", "coworkers", "supervisor",
  "mental_health_interview", "phys_health_interview",
  "mental_vs_physical", "obs_consequence", "work_interfere",
  "no_employees"
)

survey[categorical_vars] <- lapply(survey[categorical_vars], factor)

#converting mental-health varibales to numeric

#treatment
survey$treatment_num <- ifelse(survey$treatment == "Yes", 1, 0)

#work interefrence
survey$work_num <- as.numeric(survey$work_interfere) - 1

#consequences
survey$mh_conseq_num <- recode(survey$mental_health_consequence,
                                 "No" = 0, "Maybe" = 1, "Yes" = 2)

survey$ph_conseq_num <- recode(survey$phys_health_consequence,
                                 "No" = 0, "Maybe" = 1, "Yes" = 2)

#comfort in interviews
survey$interview_num <- recode(survey$mental_health_interview,
                                 "No" = 0, "Maybe" = 1, "Yes" = 2)

#observed consequenses
survey$obs_num <- recode(survey$obs_consequence,
                           "No" = 0, "Yes" = 1)

#creating the Mental Health Risk Score

survey$risk_score <-
  as.numeric(survey$treatment) +
  as.numeric(survey$phys_health_interview) +
  as.numeric(survey$seek_help) +
  as.numeric(survey$anonymity) +
  as.numeric(survey$care_options) +

```

```

as.numeric(survey$leave) +
as.numeric(survey$mental_vs_physical)

head(survey)

##   Age Gender      Country family_history treatment work_interfere
## 1 37 Female United States           No      Yes        Often
## 2 44   Male United States           No       No    Rarely
## 3 32   Male     Canada           No       No    Rarely
## 4 31   Male United Kingdom         Yes      Yes        Often
## 5 31   Male United States          No       No     Never
## 6 33   Male United States          Yes      No    Sometimes
##   no_employees remote_work tech_company   benefits care_options
## 1            6-25      No       Yes      Yes Not sure
## 2 More than 1000      No      No Don't know      No
## 3            6-25      No       Yes      No      No
## 4            26-100     No       Yes      No      Yes
## 5           100-500     Yes      Yes      Yes      No
## 6            6-25      No       Yes      Yes Not sure
##   wellness_program seek_help anonymity      leave
## 1             No      Yes      Yes Somewhat easy
## 2 Don't know Don't know Don't know      Don't know
## 3             No      No Don't know Somewhat difficult
## 4             No      No      No Somewhat difficult
## 5 Don't know Don't know Don't know      Don't know
## 6             No Don't know Don't know      Don't know
##   mental_health_consequence phys_health_consequence   coworkers supervisor
## 1             No           No Some of them      Yes
## 2           Maybe           No           No      No
## 3             No           No           Yes      Yes
## 4             Yes           Yes Some of them      No
## 5             No           No Some of them      Yes
## 6             No           No           Yes      Yes
##   mental_health_interview phys_health_interview mental_vs_physical
## 1             No           Maybe      Yes
## 2             No           No Don't know
## 3             Yes           Yes      No
## 4           Maybe           Maybe      No
## 5             Yes           Yes Don't know
## 6             No           Maybe Don't know
##   obs_consequence treatment_num work_num mh_conseq_num ph_conseq_num
## 1             No           1       1       0       0
## 2             No           0       2       1       0
## 3             No           0       2       0       0
## 4             Yes          1       1       2       2
## 5             No           0       0       0       0
## 6             No           0       3       0       0
##   interview_num obs_num risk_score
## 1            0      0      17
## 2            0      0       8
## 3            2      0      12
## 4            1      1      14
## 5            2      0       9
## 6            0      0       8

```

For Data Preprocessing I did the following:

1. Checked the dataset for missing values and found that some columns, like comments, had over 1,000 missing entries, and others like state and self_employed had very high missingness. I removed them because they wouldn't be useful for analysis.
2. Handled missing values in important variables. For work_interfere, which is closely related to mental health and had many missing entries, I kept the data by creating a new category called "Unknown" instead of deleting those rows.
3. Cleaned the Gender column, which had many inconsistent entries (like "male", "M", "Man", "f", "Female", "cis female", etc.). I standardized all gender labels into three groups: Male, Female, and Other so the variable could be analyzed properly.
4. Converted several character variables into factors, such as treatment, benefits, family_history, remote_work, no_employees, and others. This makes them usable in statistical modeling since they represent categories rather than text.
5. Standardized categories for variables that had inconsistent labels. I cleaned and standardized several mental-health variables that had inconsistent labels, such as mental_health_consequence, phys_health_consequence, mental_health_interview, obs_consequence, work_interfere, treatment, seek_help, anonymity, care_options, leave, and mental_vs_physical. These were recoded into clear, consistent categories (e.g., No = 0, Maybe = 1, Yes = 2) so they could be accurately converted into numeric values for analysis.
6. Created numeric versions of mental-health-related variables. Questions like treatment (Yes/No), mental health consequences, physical consequences, work interference, interview comfort, and observed consequences were turned into numeric values.
7. I built the Mental Health Risk Score by combining the numeric versions of treatment, phys_health_interview, seek_help, anonymity, care_options, leave, and mental_vs_physical. These variables capture help-seeking, perceived support, and workplace stigma, so summing them creates a single continuous score representing overall mental-health risk.
8. Removed irrelevant variables, like Timestamp, which don't contribute to mental-health prediction and only add extra noise.

Predictor Variables I chose and why:

Age: Included because mental-health experiences can vary across life stages, though effects may be small.

Gender: Gender can influence help-seeking and stigma perceptions, making it an important demographic factor.

Family history: Family background is a well-known risk factor for mental-health vulnerability.

Supervisor support: Supervisor behavior strongly affects workplace stress and comfort seeking help.

Remote work: Work setting can change stress and isolation levels.

Company size (no_employees): Organizational size affects workplace culture and availability of support resources.

Work interference: One of the strongest indicators of mental-health strain; directly measures functional impact.

Tech company: Captures cultural differences unique to the tech industry.

Benefits: Access to mental-health benefits can influence support-seeking behavior and perceived stability.

Coworker support: Social environment plays a key role in stress and mental-health burden.

Physical health consequences: Measures whether the workplace stigmatizes physical conditions, helping contrast attitudes.

Mental-health consequences: Captures perceived workplace stigma and fear of negative outcomes.

#Model Creation

```
# Linear Regression Model
model <- lm(
  risk_score ~ Age +
    Gender +
    family_history +
    supervisor +
    remote_work +
    no_employees +
    work_interfere +
    tech_company +
    benefits +
    coworkers +
    phys_health_consequence +
    mental_health_consequence,
  data = survey
)

summary(model)

##
## Call:
## lm(formula = risk_score ~ Age + Gender + family_history + supervisor +
##     remote_work + no_employees + work_interfere + tech_company +
##     benefits + coworkers + phys_health_consequence + mental_health_consequence,
##     data = survey)
##
## Residuals:
##      Min        1Q        Median         3Q        Max 
## -8.1524 -2.0242 -0.0506  1.7742  7.8367 
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)    
## (Intercept)           10.158019  0.6016411 16.884 < 2e-16 ***
## Age                  0.0001965  0.0112870  0.017 0.986114    
## GenderMale            -0.2156935  0.2047735 -1.053 0.292399    
## GenderOther            1.1250748  0.8957528  1.256 0.209352    
## family_historyYes     0.1851490  0.1741954  1.063 0.288045    
## supervisorSome of them 0.4241999  0.2296163  1.847 0.064926 .  
## supervisorYes          1.1494988  0.2565940  4.480 8.17e-06 ***
## remote_workYes         0.1492911  0.1828032  0.817 0.414272    
## no_employees100-500   -1.3089124  0.3294766 -3.973 7.52e-05 *** 
## no_employees26-100    -1.1344728  0.2949066 -3.847 0.000126 *** 
## no_employees500-1000  -0.7997138  0.4499405 -1.777 0.075754 .  
## no_employees6-25       -1.0213779  0.2847410 -3.587 0.000348 *** 
## no_employeesMore than 1000 -0.7410394  0.3238658 -2.288 0.022301 *  
## work_interfereOften    1.5920028  0.3152970  5.049 5.11e-07 *** 
## work_interfereRarely   1.0008194  0.2914953  3.433 0.000616 *** 
## work_interfereSometimes 1.3636672  0.2416883  5.642 2.08e-08 *** 
## work_interfereUnknown   0.0189611  0.2572231  0.074 0.941250
```

```

## tech_companyYes      -0.4345328  0.2126822  -2.043 0.041255 *
## benefitsNo          1.9869936  0.2139582   9.287 < 2e-16 ***
## benefitsYes         2.8824433  0.1986685  14.509 < 2e-16 ***
## coworkersSome of them 0.1233654  0.2306553   0.535 0.592852
## coworkersYes        0.5440593  0.3160496   1.721 0.085424 .
## phys_health_consequenceNo 0.2949602  0.2147365   1.374 0.169820
## phys_health_consequenceYes 1.1322254  0.4176042   2.711 0.006797 **
## mental_health_consequenceNo 1.2396300  0.2120704   5.845 6.47e-09 ***
## mental_health_consequenceYes 0.2070718  0.2370063   0.874 0.382454
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.755 on 1225 degrees of freedom
## Multiple R-squared:  0.3146, Adjusted R-squared:  0.3007
## F-statistic:  22.5 on 25 and 1225 DF,  p-value: < 2.2e-16
install.packages("Metrics")

## Installing package into '/cloud/lib/x86_64-pc-linux-gnu-library/4.5'
## (as 'lib' is unspecified)
library(Metrics)

# Predictions from the model
pred <- predict(model, survey)

# R-squared
rsq <- summary(model)$r.squared

# Mean Squared Error
mse_val <- mse(survey$risk_score, pred)

# Mean Absolute Error
mae_val <- mae(survey$risk_score, pred)

rsq

## [1] 0.3146436
mse_val

## [1] 7.434327
mae_val

## [1] 2.202786

```

Model Performance Interpretation

The linear regression model showed moderate performance. It achieved an R-squared of 0.3146, meaning the model explains about 31% of the variation in mental-health risk scores. While this is not extremely high, it is probably because real-world survey data contains many subjective and categorical responses. The Mean Squared Error (MSE) was 7.43, and the Root Mean Squared Error (RMSE) was 2.20, which means that on average, the model's predictions differ from the actual risk score by about two points. This level of error is expected with self-reported mental-health data.

The coefficients suggest that workplace environment factors, especially work interference, benefits, mental

health consequences, and supervisor support, have the strongest influence on the risk score. In contrast, demographic factors such as age and gender contribute much less

Interpretation of Model Coefficients

Age: The effect of age is very small and not statistically significant, meaning age does not meaningfully change mental health risk in this dataset.

Gender: Gender differences are small and not statistically significant, so mental health risk does not strongly differ between men, women, or participants who selected “other.”

Family history: Participants with a family history of mental illness have slightly higher predicted risk scores, although the effect becomes weaker when workplace factors are included.

Supervisor support: Participants who reported having supportive supervisors show higher predicted risk scores. This is likely because individuals experiencing more symptoms often seek support, so the model captures an association rather than a causal relationship.

Remote work: Remote work has a very small and non significant effect. Working remotely does not meaningfully change mental health risk in this dataset.

Company size (number of employees): Employees at smaller companies show lower predicted risk scores compared to those in very large companies. This suggests that mental health burdens may be reported more often in larger corporate environments.

Work interference: Work interference is one of the strongest predictors in the model. Participants whose mental health “Rarely,” “Sometimes,” or “Often” interferes with their work have much higher predicted risk scores compared to those who selected “Never.” Greater interference clearly corresponds to higher mental health risk.

Tech company: Working at a tech company slightly decreases the risk score, although the effect is small.

Benefits: Participants who reported having benefits or not having benefits both show higher predicted risk scores compared to those who answered “Don’t know.” This pattern appears because the risk score is based on help seeking and awareness related variables. People who experience more symptoms tend to be more aware of their benefits.

Coworkers: Reporting supportive coworkers has a small positive effect but it is not statistically significant.

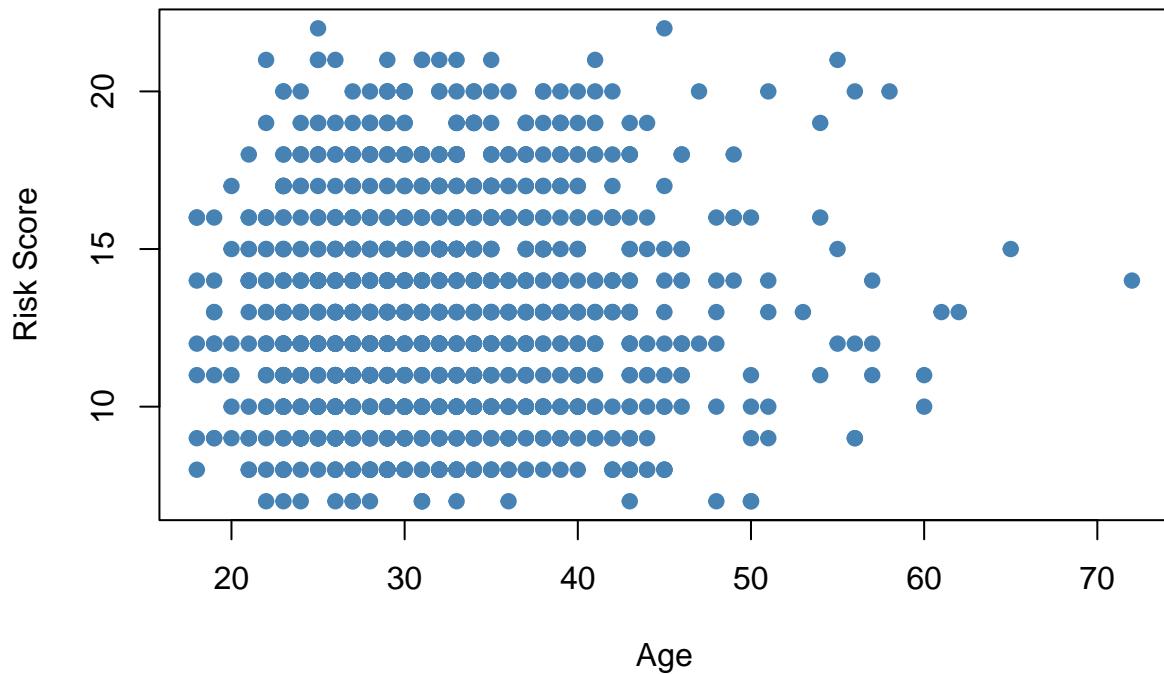
Physical health consequence: Participants who reported seeing negative physical health consequences at their workplace have higher predicted mental health risk scores. This suggests that environments where negative outcomes are visible may increase stress or perceived risk.

Mental health consequence: Participants who reported negative consequences for discussing mental health have significantly higher predicted risk scores. Workplace stigma appears to be a major factor increasing mental health burden.

Visualisation and Analysis

```
#Scatterplot: Age vs Risk Score
plot(survey$Age, survey$risk_score,
      main="Risk Score vs Age",
      xlab="Age", ylab="Risk Score",
      pch=19, col="steelblue")
```

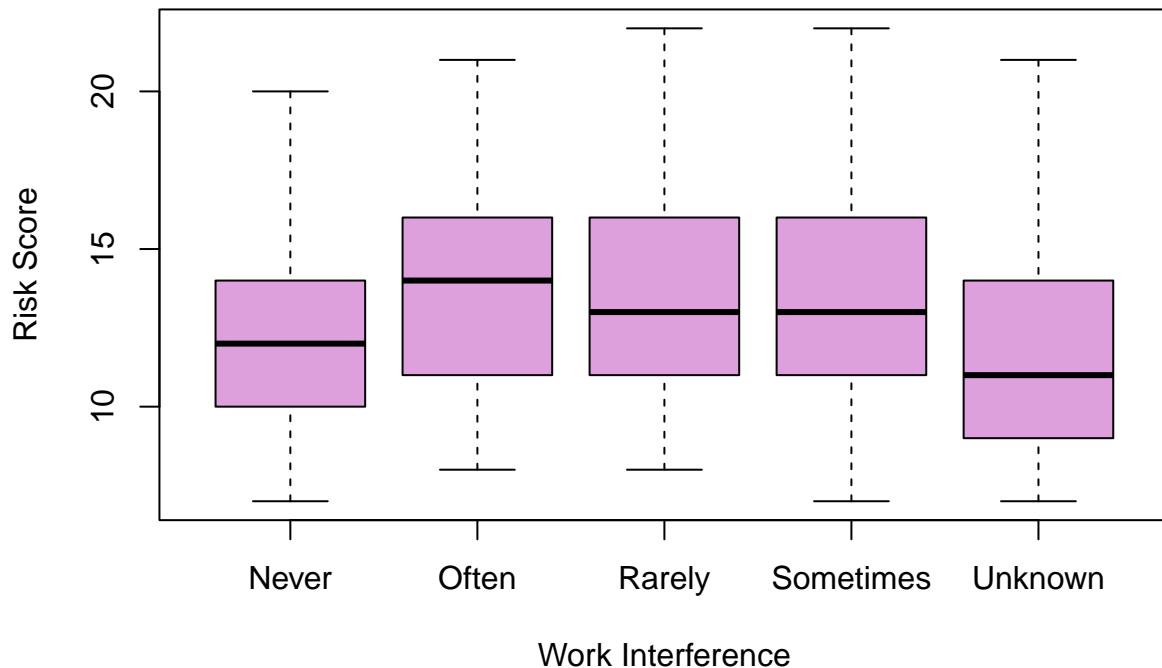
Risk Score vs Age



The scatter plot shows no strong pattern between age and risk score, which matches the regression results. Mental-health risk appears relatively consistent across age groups.

```
#Boxplot: Risk Score by Work Interference
boxplot(risk_score ~ work_interfere, data = survey,
        main = "Risk Score by Work Interference Level",
        xlab = "Work Interference",
        ylab = "Risk Score",
        col = "plum")
```

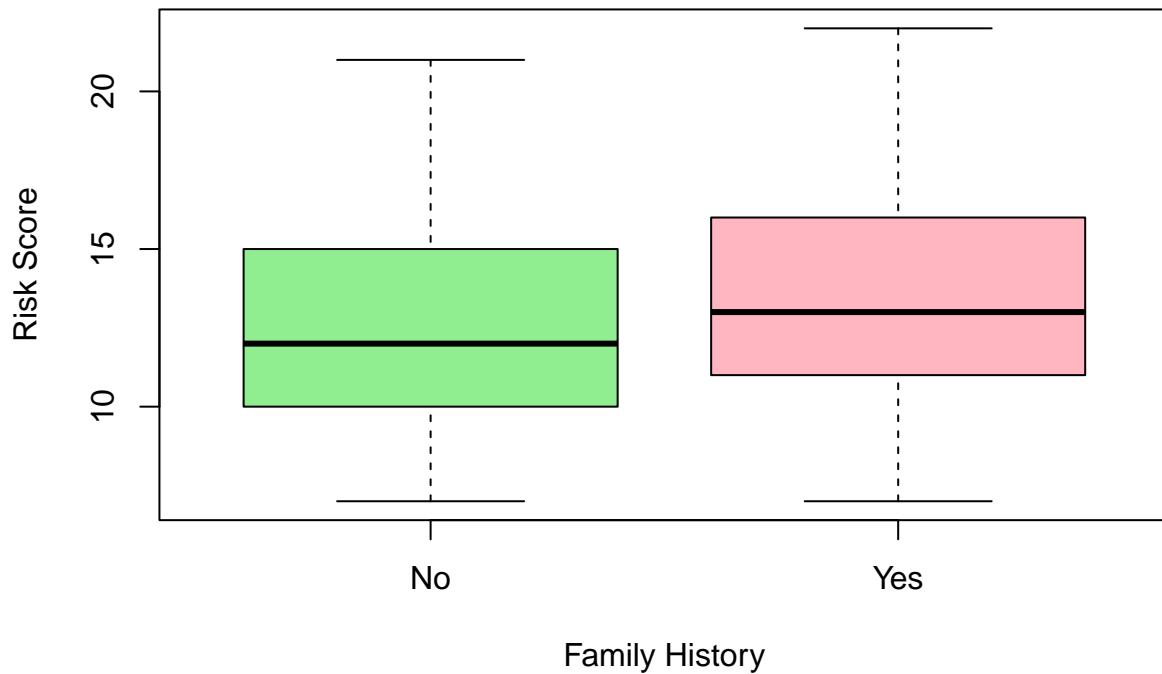
Risk Score by Work Interference Level



Risk scores increase steadily as work interference worsens from “Never” to “Often.” This makes work interference one of the strongest behavioral indicators of mental-health strain.

```
#Boxplot: Risk Score by Family History
boxplot(risk_score ~ family_history, data = survey,
        main = "Risk Score by Family History",
        xlab = "Family History",
        ylab = "Risk Score",
        col = c("lightgreen", "lightpink"))
```

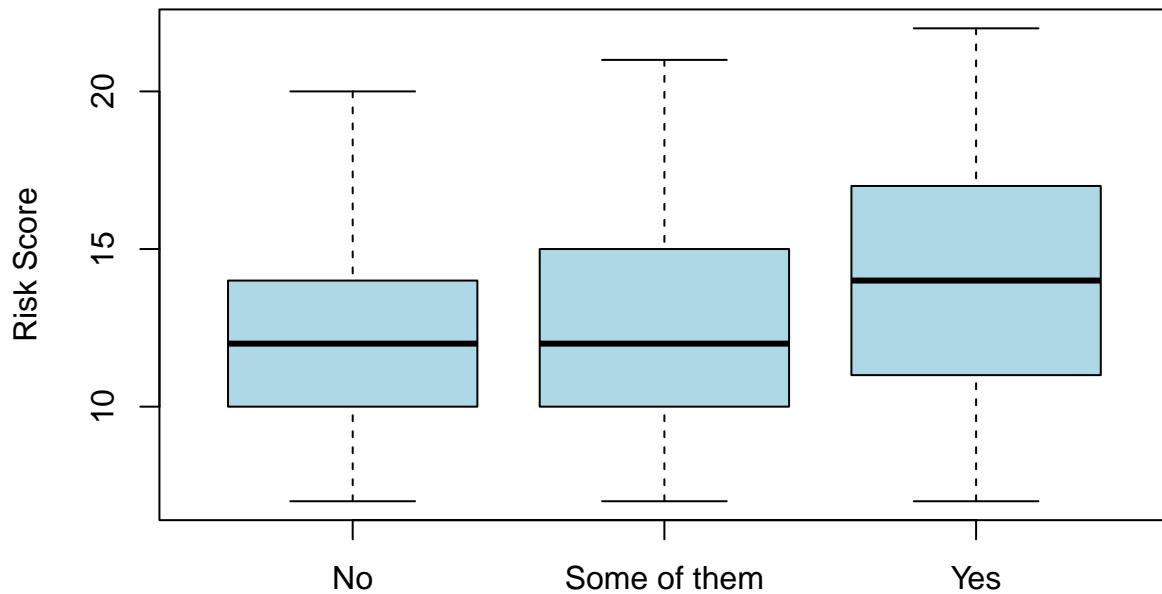
Risk Score by Family History



Participants with a family history of mental illness have higher risk scores on average. This confirms that family background is an important mental-health risk factor.

```
#Boxplot: Risk Score by Supervisor Support
boxplot(risk_score ~ supervisor, data = survey,
        main = "Risk Score by Supervisor Support",
        xlab = "Supervisor Support",
        ylab = "Risk Score",
        col = "lightblue")
```

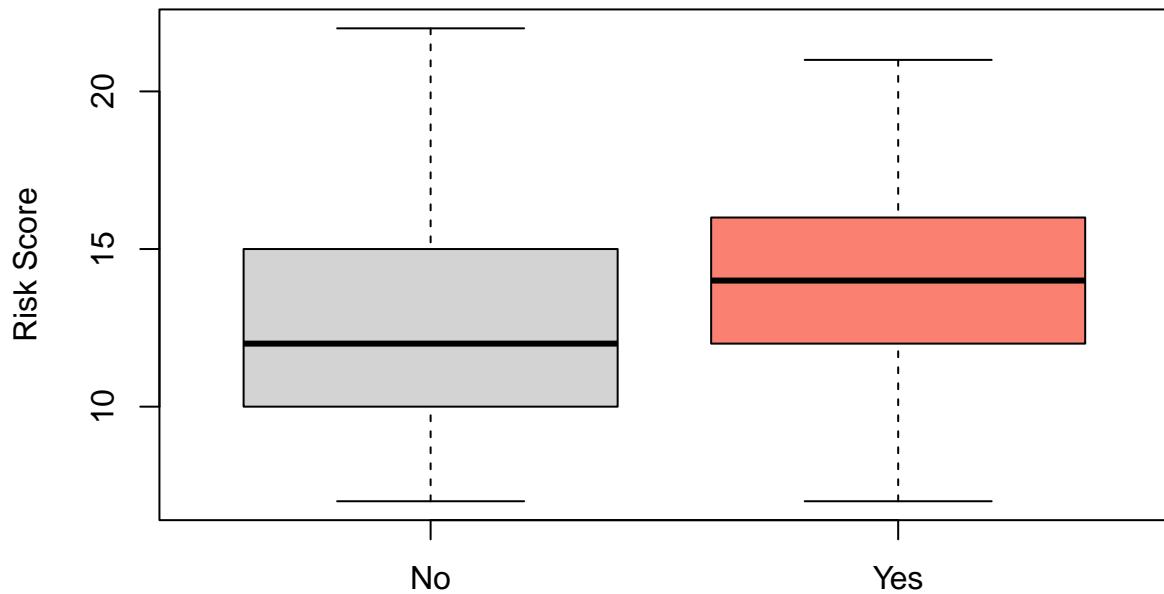
Risk Score by Supervisor Support



Employees who reported having supportive supervisors ('Yes' or 'Some of them') actually showed higher mental-health risk scores than those with no supervisor support. This may indicate that people with more mental-health challenges are more likely to seek out or notice supervisor support

```
#Boxplot: Risk Score by Observed Consequences
boxplot(risk_score ~ obs_consequence, data = survey,
        main = "Risk Score by Observed Workplace Consequences",
        xlab = "Observed Consequences",
        ylab = "Risk Score",
        col = c("lightgrey", "salmon"))
```

Risk Score by Observed Workplace Consequences



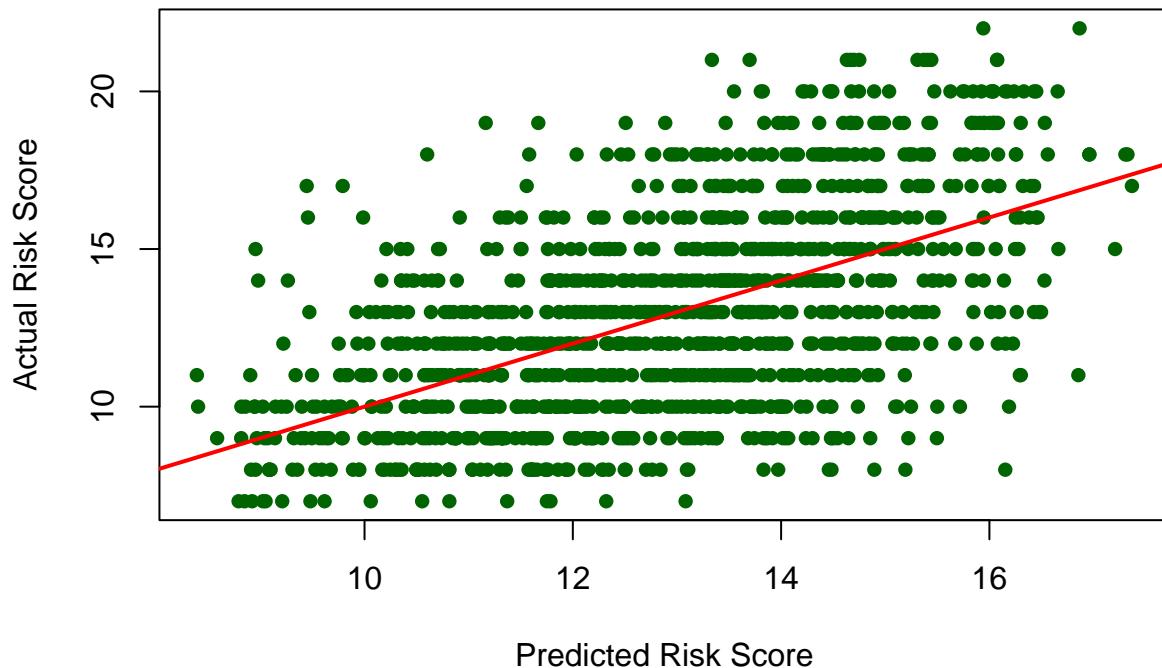
Observed Consequences

Peo-

ple who reported seeing negative consequences for coworkers have higher median risk scores. This supports the idea that workplace stigma and negative experiences elevate mental-health risk.

```
# Actual vs Predicted Plot
plot(pred, survey$risk_score,
      main = "Actual vs Predicted Risk Score",
      xlab = "Predicted Risk Score",
      ylab = "Actual Risk Score",
      pch = 16, col = "darkgreen")
abline(0, 1, col = "red", lwd = 2)
```

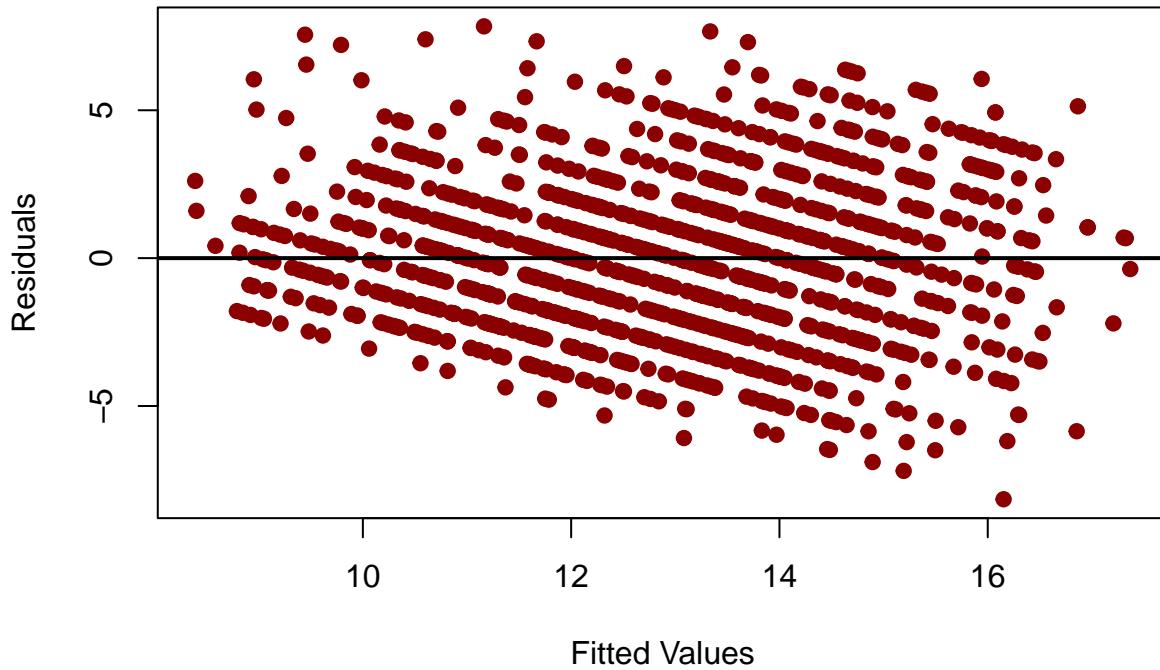
Actual vs Predicted Risk Score



There is a positive trend between actual and predicted values, meaning the model captures the general direction of the relationship. But the large vertical spread shows the predictions are not very precise.

```
#Residual Plot
plot(model$fitted.values, resid(model),
     main="Residuals vs Fitted Values",
     xlab="Fitted Values", ylab="Residuals",
     pch=19, col="darkred")
abline(h=0, lwd=2)
```

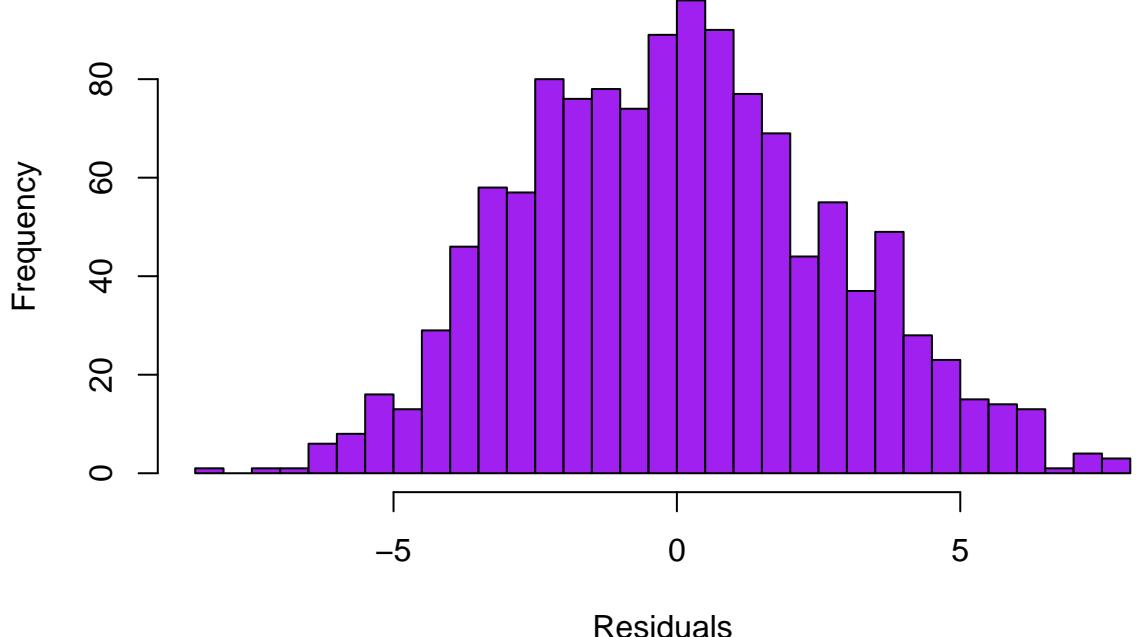
Residuals vs Fitted Values



The residuals show a clear funnel-shaped pattern rather than random scatter, meaning the model violates homoscedasticity. This suggests the linear model may not fully capture the true relationships in the data.

```
#Histogram: Residuals
hist(resid(model), breaks=30,
     main="Residual Distribution",
     col="purple", xlab="Residuals")
```

Residual Distribution



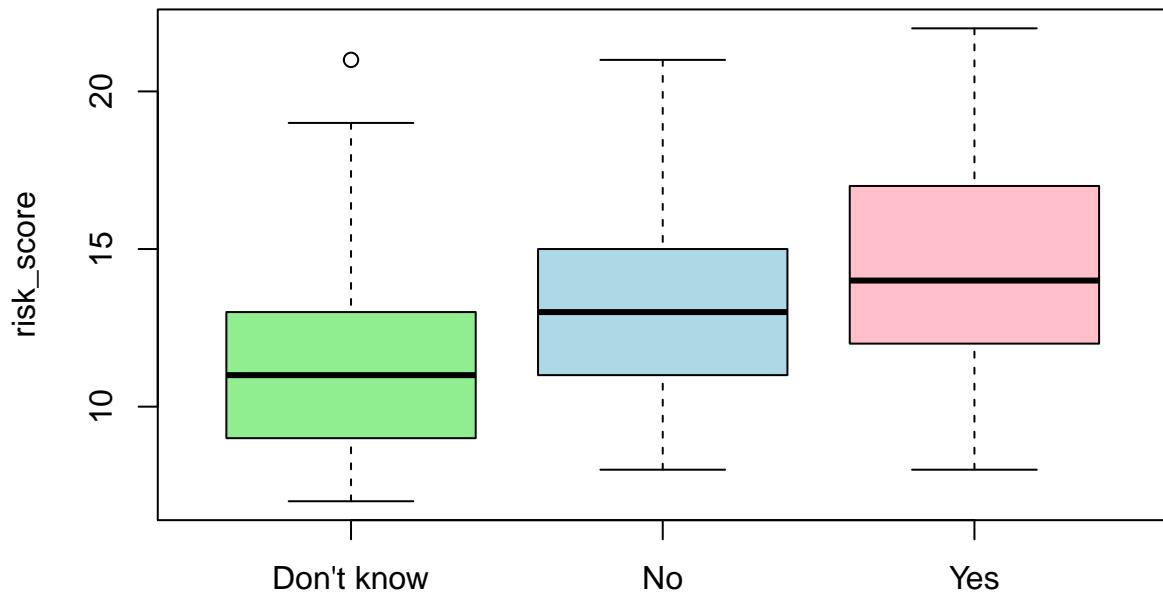
Residuals

This his-

togram shows that the residuals are roughly centered around zero, which suggests the model does not consistently over- or under-predict. However, the spread is wide, indicating substantial prediction error.

```
# Boxplot: Risk Score by Benefits
boxplot(risk_score ~ benefits, data=survey,
        main="Risk Score by Benefits",
        col=c("lightgreen","lightblue","pink"))
```

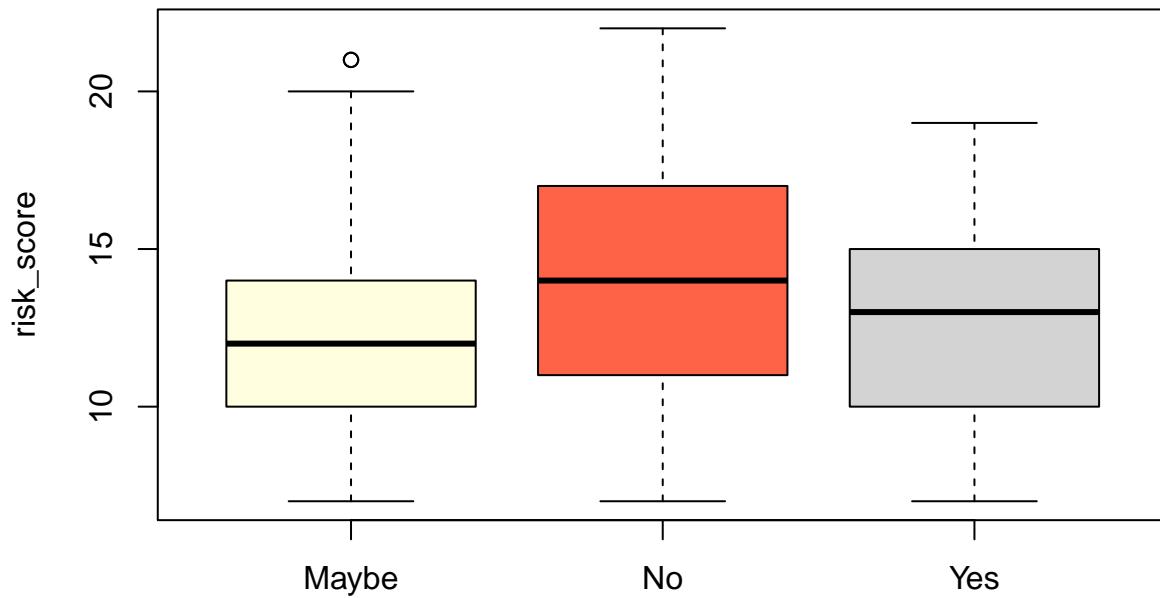
Risk Score by Benefits



People who answered “Don’t know” about benefits have the lowest risk scores. Participants who answered “Yes” or “No” show higher scores, likely because individuals experiencing symptoms are more aware of their benefits and policies.

```
# Boxplot: Risk Score by Mental Health Consequence
boxplot(risk_score ~ mental_health_consequence, data=survey,
        main="Risk Score by Mental Health Consequence",
        col=c("lightyellow","tomato","lightgray"))
```

Risk Score by Mental Health Consequence



mental_health_consequence

Participants who answered “No” tend to have the highest mental-health risk scores, while those who answered “Yes” show slightly lower scores. This suggests that people who believe mental-health issues would not lead to consequences may actually report more symptoms, possibly because they feel safer disclosing them.