Modeling Workplace Stress and Health Behaviors Using R

```
# Load data
data <- read.csv("Work.csv")

# Check structure
str(data)
```

'data.frame': 1200 obs. of 12 variables:

$ X : int 1 2 3 4 5 6 7 8 9 10 ...

$ age : num 46.9 35 33.5 43.8 21 49 35.9 40.7 52.9 31 ...

$ income : int 20981 86007 62126 36758 54404 27894 62511 25915 61366 31358 ...

$ hours_worked : num 37.2 42 39.5 34.6 45.5 42.6 42.1 34 40.1 40.3 ...

$ stress_level : num 50.1 47.3 50.7 49.2 51.2 56.6 44.4 49.8 38.4 48.8 ...

$ gender : chr "Male" "Female" "Female" "Male" ...

$ smoker : chr "Yes" "No" "No" "No" ...

$ education : chr "High School" "College" "High School" "High School" ...

$ job_sector : chr "Education" "Education" "Retail" "Health" ...

$ doctor_visits: int 0 0 0 0 1 3 1 0 1 0 ...

$ sick_days : int 1 0 0 1 1 0 1 1 3 0 ...

Descriptive Statistics

```
# Numeric summary for age
summary(data$age)
```

Min. 1st Qu. Median Mean 3rd Qu. Max.

2.30 27.70 34.60 34.81 41.50 72.80

> #Comment:

> #The ages of individuals in the dataset range from 2.3 to 72.8 years, with an average (mean) of 34.81 years.

> #The median age is 34.6, indicating a fairly symmetric distribution.

> #The majority of respondents fall within the interquartile range of 27.7 to 41.5 years, suggesting the data is concentrated among working-age adults

>

```
> # integer summary for hours_worked(int)

> summary(data$hours_worked)
```

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 20.60   36.70   40.10   40.05   43.50   54.20
```

>

> #Comment:

> #The number of hours worked per week ranges from 20.6 to 54.2 hours, with a mean of 40.05 hours and a median of 40.1 hours.

> #This suggests a consistent full-time work schedule across most individuals.

> #The distribution appears to be symmetric, with 50% of individuals working between 36.7 and 43.5 hours per week

>

```
> # Frequency for gender(char)

> table(data$gender)
```

```
Female   Male
  610    590
```

> #Comment:

> #The gender distribution is nearly balanced, with 610 females and 590 males in the dataset.

> #This even split ensures that gender-based comparisons are likely to be statistically meaningful and not biased by unequal sample sizes.

Random Sample of 250 Observations

```
set.seed(123)
sample_data <- data[sample(nrow(data), 250), ]
head(sample_data, 10)
```

	X	age	income	hours_worked	stress_level	gender	smoker	education	job_sector
355	355	40.2	65978	44.3	51.9	Male	No	Graduate	Retail
613	613	25.6	72820	36.4	45.0	Female	Yes	Graduate	Health
847	847	38.1	67264	39.6	40.1	Male	Yes	Graduate	Tech
1163	1163	33.9	22373	41.0	64.1	Female	No	College	Tech
489	489	41.0	41368	37.0	38.6	Male	No	High School	Retail
1167	1167	33.7	61434	37.2	48.1	Female	No	College	Health
1140	1140	46.9	55875	30.3	58.6	Male	Yes	Graduate	Education
954	954	23.3	43076	28.0	60.6	Male	No	College	Retail
42	42	26.9	48623	51.1	33.3	Female	Yes	High School	Tech
126	126	20.0	49756	33.5	36.0	Male	No	High School	Health

	doctor_visits	sick_days	high_stress
355	0	0	Yes
613	0	0	Yes
847	1	0	No
1163	1	0	No
489	0	0	Yes
1167	1	0	No
1140	1	0	Yes
954	0	0	No
42	0	0	Yes
126	0	0	No

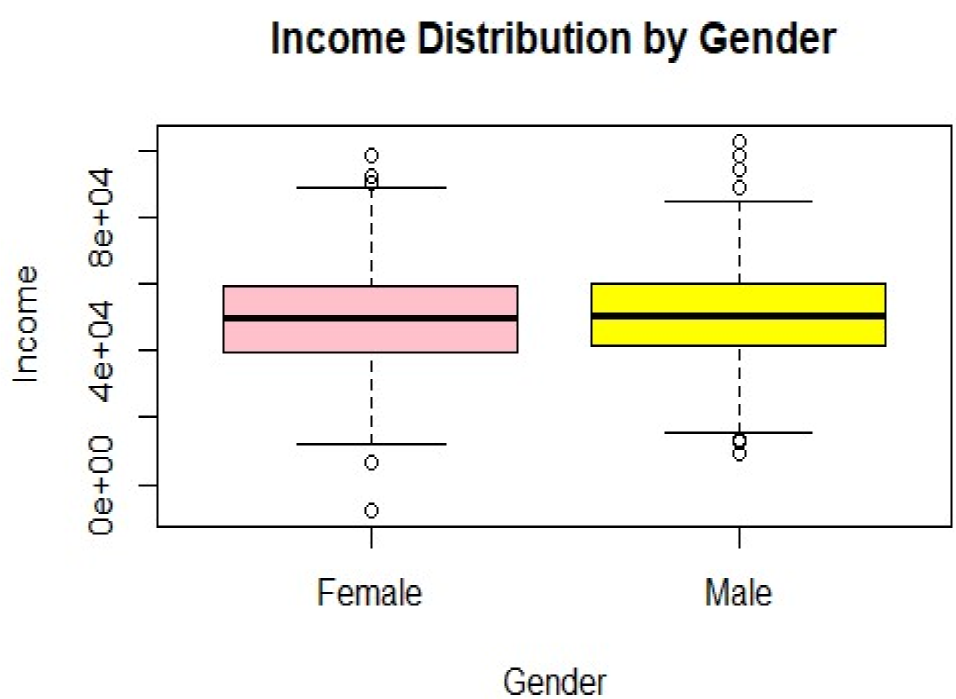
> #Comment:

> #The gender distribution is nearly balanced, with 610 females and 590 males in the dataset.

> #This even split ensures that gender-based comparisons are likely to be statistically meaningful and not biased by unequal sample sizes.

Boxplot of Income by Gender

```
boxplot(income ~ gender, data = data,
        main = "Income Distribution by Gender",
        xlab = "Gender", ylab = "Income",
        col = c("pink", "yellow"))
```



t-test for Income by Gender

> #Assumptions Made:

> #1. The gender variable is binary (Male/Female).

> #2. The income variable is quantitative and continuous.

> #3. The two groups (Male and Female) are independent.

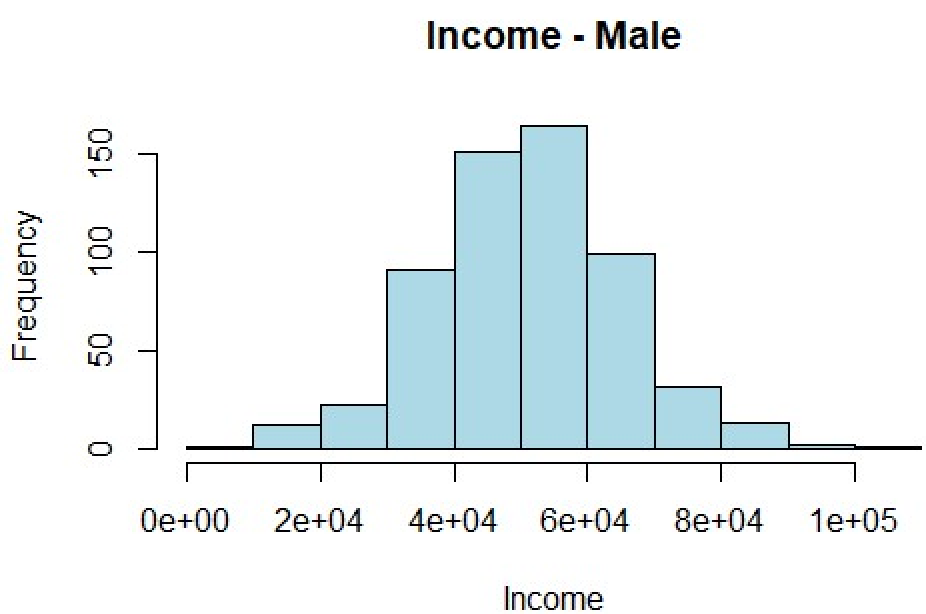
> #4. Each group's income is approximately normally distributed.

> #5. Variance between the two groups may or may not be equal

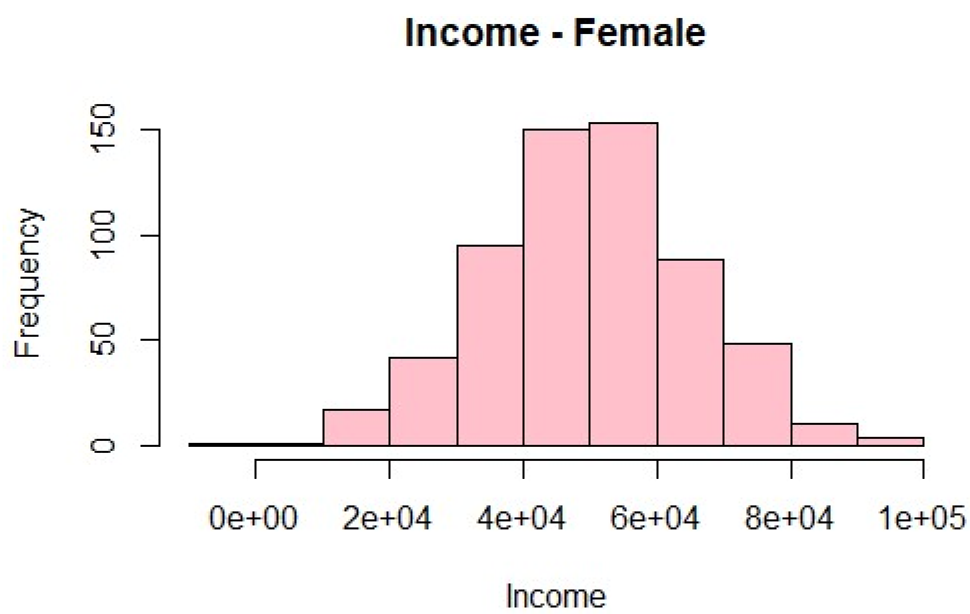
> # Normality Check

> # Histogram for each gender

```
> hist(data$income[data$gender == "Male"], main = "Income - Male", xlab
= "Income", col = "lightblue")
```




```
> hist(data$income[data$gender == "Female"], main = "Income - Female",
xlab = "Income", col = "pink")
```



> # Shapiro-Wilk test for normality

```
> shapiro.test(data$income[data$gender == "Male"])
```

Shapiro-Wilk normality test

data: data$income[data$gender == "Male"]

W = 0.99663, p-value = 0.2523

```
> shapiro.test(data$income[data$gender == "Female"])
```

Shapiro-Wilk normality test

data: data$income[data$gender == "Female"]

W = 0.99833, p-value = 0.8313

>

> #Interpretation:

> #Since the p-value = 0.8313 > 0.05, we fail to reject the null hypothesis of normality.

> #This means the income data for females is approximately normally distributed.

>

> #Test for Equal Variances

> # F-test for variance

```
> var.test(income ~ gender, data = data)
```

F test to compare two variances

data: income by gender

F = 1.2165, num df = 609, denom df = 589, p-value = 0.0167

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

1.036165 1.427899

sample estimates:

ratio of variances

1.216496

> #Interpretation:

> #Since the p-value = 0.0167 < 0.05, we reject the null hypothesis.

> #This means the variances are significantly different across gender groups.

> #Hypothesis Testing - t-test

```
> male_income <- data$income[data$gender == "Male"]

> female_income <- data$income[data$gender == "Female"]

> t.test(male_income, female_income, var.equal = FALSE)
```

Welch Two Sample t-test

data: male_income and female_income

t = 1.0872, df = 1193, p-value = 0.2772

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

-754.5215 2630.1436

sample estimates:

mean of x mean of y

50843.29 49905.48

Logistic Regression for High Stress

```
data$high_stress_bin <- ifelse(data$high_stress == "Yes", 1, 0)
data$smoker_bin <- ifelse(data$smoker == "Yes", 1, 0)
model_logit <- glm(high_stress_bin ~ age + income + hours_worked +
smoker_bin,
                   data = data, family = binomial)
summary(model_logit)
```

Call:

glm(formula = high_stress_bin ~ age + income + hours_worked +

smoker_bin, family = binomial, data = data)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -4.720e+00 5.835e-01 -8.089 6.0e-16 ***

age 3.450e-04 5.821e-03 0.059 0.95273

income 5.158e-06 4.025e-06 1.281 0.20003 hours_worked 9.940e-02 1.267e-02   7.843 4.4e-15 ***

smoker_bin   3.546e-01 1.205e-01   2.942 0.00326 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 1642.2 on 1199 degrees of freedom

Residual deviance: 1564.7 on 1195 degrees of freedom

AIC: 1574.7

Number of Fisher Scoring iterations: 4

Interpretation of Coefficients

#The model suggests that hours worked per week and smoking status are statistically significant predictors of high stress.

> #Specifically:

> #A 1-hour increase in work hours is associated with a log-odds increase of 0.0994 in being highly stressed, holding other variables constant.

> #Smokers have higher odds of experiencing high stress than non-smokers.

> #Age and income were not significant predictors (p > 0.05), indicating they do not have a meaningful effect in this model.

Confusion Matrix

```
pred_probs <- predict(model_logit, type = "response")
pred_class <- ifelse(pred_probs > 0.5, 1, 0)
table(Predicted = pred_class, Actual = data$high_stress_bin)
```

Actual

Predicted   0   1

0 530 315

1 150 205

Linear Regression for Sick Days

```
model_lm <- lm(sick_days ~ age + income + stress_level + education +
smoker, data = data)
summary(model_lm)
```

```
Call:
lm(formula = sick_days ~ age + income + stress_level + education +
    smoker, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9922 -0.4951 -0.3211  0.4495  4.3288

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -2.735e-01  1.500e-01  -1.823   0.0685 .
age                 4.196e-03  2.035e-03   2.062   0.0394 *
income             -9.113e-07  1.407e-06  -0.648   0.5173
stress_level        1.180e-02  2.050e-03   5.756 1.09e-08 ***
educationGraduate  -4.121e-02  5.178e-02  -0.796   0.4263
educationHigh School -1.012e-01  5.214e-02  -1.942   0.0524 .
smokerYes           2.182e-01  4.207e-02   5.186 2.52e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7274 on 1193 degrees of freedom
```

Multiple R-squared: 0.05425, Adjusted R-squared: 0.04949

F-statistic: 11.41 on 6 and 1193 DF, p-value: 1.983e-12

Interpretation of Coefficients

Predictor	**Coefficient**	**p-value**	**Interpretation**
stress_level	**0.0118**	**< 0.001**	**Significant**: For every 1-unit increase in stress score, sick days increase by about 0.012 days.
smokerYes	**0.2182**	**< 0.001**	**Significant**: Smokers take about **0.22 more sick days** per year than non-smokers, on average.

The model shows that individuals with **higher stress levels** and those who **smoke** are more likely to take more sick days. These two variables are statistically significant predictors and have practical implications for health and workplace interventions.