Remote Work and Mental Health

Amy Fishencord

2024-10-14

Loading in the Data

First we will load in our data using the read.csv() function

```
knitr::opts_chunk$set(echo = TRUE)
df = read.csv("/Users/amyfishencord/Downloads/Impact_of_Remote_Work_on_Mental_Health.csv",
              header=TRUE)
```

Importing Libraries

Next, we will be importing libraries used to produce data visualization and complete data manipulation.

```
library(ggplot2)
library(dplyr)
library(plotrix)
library(plotly)
library(knitr)
library(naniar)
library(RColorBrewer)
```

About the Data

The "Remote Work and Mental Health" dataset explores the effects of remote work on employees' mental well-being. It includes 5,000 records collected from employees world-wide that capture various factors such as stress levels, job satisfaction, and feelings of social isolation among workers across different industries and job roles.

Quick Overview of the Data

I will be using the str() function to show each column name and the first few values in the dataset to get a quick overview of the data and datatypes we will be using.

```
## 'data.frame':    5000 obs. of  20 variables:
##  $ Employee_ID                      : chr  "EMP0001" "EMP0002" "EMP0003" "EMP0004" ...
##  $ Age                              : int  32 40 59 27 49 59 31 42 56 30 ...
##  $ Gender                           : chr  "Non-binary" "Female" "Non-binary" "Male" ...
##  $ Job_Role                         : chr  "HR" "Data Scientist" "Software Engineer" "Software Engin
##  $ Industry                         : chr  "Healthcare" "IT" "Education" "Finance" ...
##  $ Years_of_Experience              : int  13 3 22 20 32 31 24 6 9 28 ...
##  $ Work_Location                    : chr  "Hybrid" "Remote" "Hybrid" "Onsite" ...
##  $ Hours_Worked_Per_Week            : int  47 52 46 32 35 39 51 54 24 57 ...
##  $ Number_of_Virtual_Meetings       : int  7 4 11 8 12 3 7 7 4 6 ...
##  $ Work_Life_Balance_Rating         : int  2 1 5 4 2 4 3 3 2 1 ...
##  $ Stress_Level                     : chr  "Medium" "Medium" "Medium" "High" ...
```

```
## $ Mental_Health_Condition          : chr  "Depression" "Anxiety" "Anxiety" "Depression" ...
## $ Access_to_Mental_Health_Resources: chr  "No" "No" "No" "Yes" ...
## $ Productivity_Change              : chr  "Decrease" "Increase" "No Change" "Increase" ...
## $ Social_Isolation_Rating          : int  1 3 4 3 3 5 5 5 2 2 ...
## $ Satisfaction_with_Remote_Work    : chr  "Unsatisfied" "Satisfied" "Unsatisfied" "Unsatisfied" ...
## $ Company_Support_for_Remote_Work  : int  1 2 5 3 3 1 3 4 4 1 ...
## $ Physical_Activity                : chr  "Weekly" "Weekly" "None" "None" ...
## $ Sleep_Quality                    : chr  "Good" "Good" "Poor" "Poor" ...
## $ Region                           : chr  "Europe" "Asia" "North America" "Europe" ...
```

Missing Values

There are no missing values within the variables of our dataset, making our dataset complete.

```r
gg_miss_var(df) +
  labs(title = "Missing Values Summary")
```

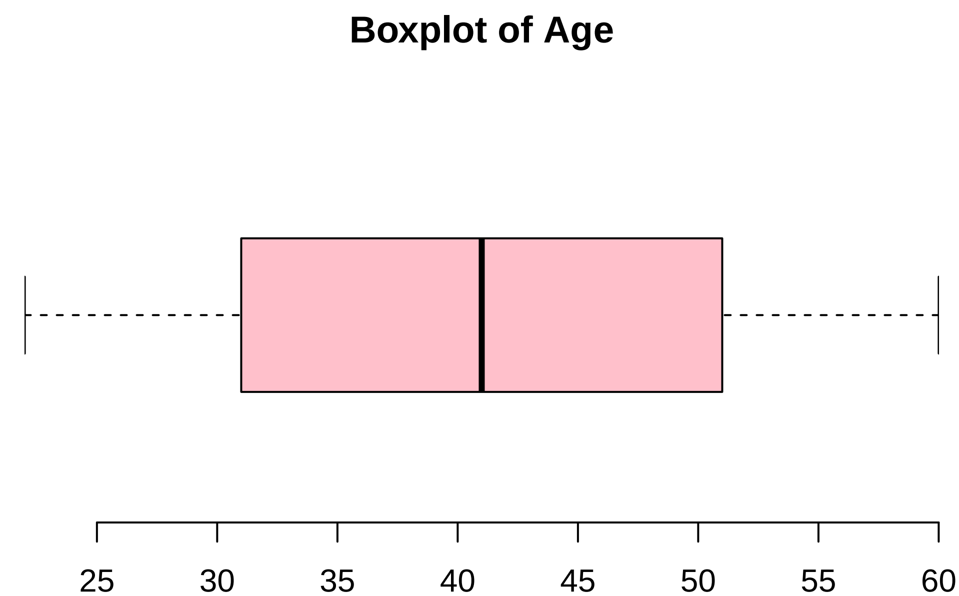
Overview of variables

I want to focus on looking at a few specific variables to get a better understanding of what they mean and what their values are. First, I want to look closer into the Age variable, specifically at the summary.

```
summary(df$Age)
```

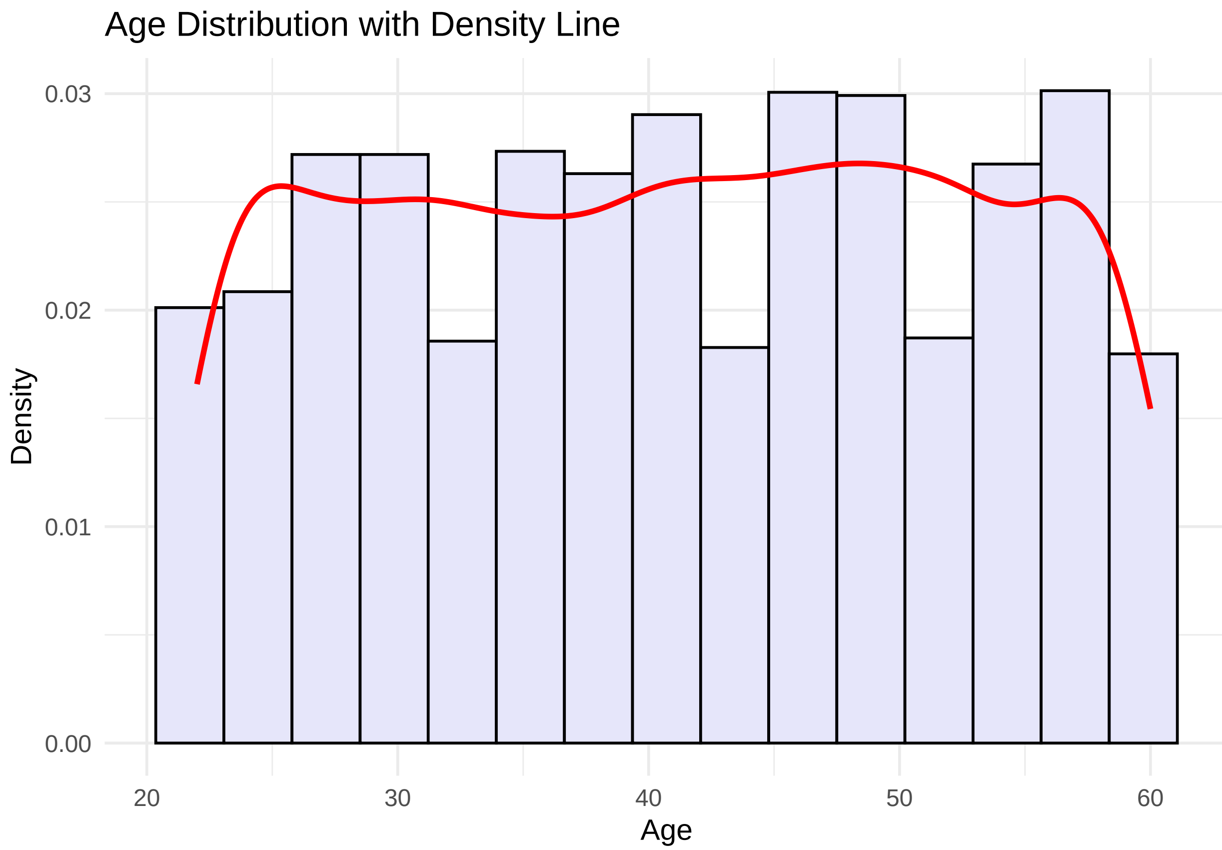
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##      22      31      41      41      51      60 
```

```
par(xaxs = "i")
boxplot(df$Age, horizontal = T, main = "Boxplot of Age", axes = F, col = "pink")
axis(1)
```



Distribution of Age

```r
suppressWarnings(
  ggplot(df, aes(x = as.numeric(Age))) +
  geom_histogram(aes(y = after_stat(density)), bins = 15, color = "black", fill = "#E6E6FA")+
  geom_density(color = "red", size = 1) +
  labs(title = "Age Distribution with Density Line", x = "Age", y = "Density") +
  theme_minimal()
)
```



The age distribution is multinomial, showing multiple peaks and valleys, which is important for understanding the diverse experiences and perspectives of employees across different age groups in the context of remote work.

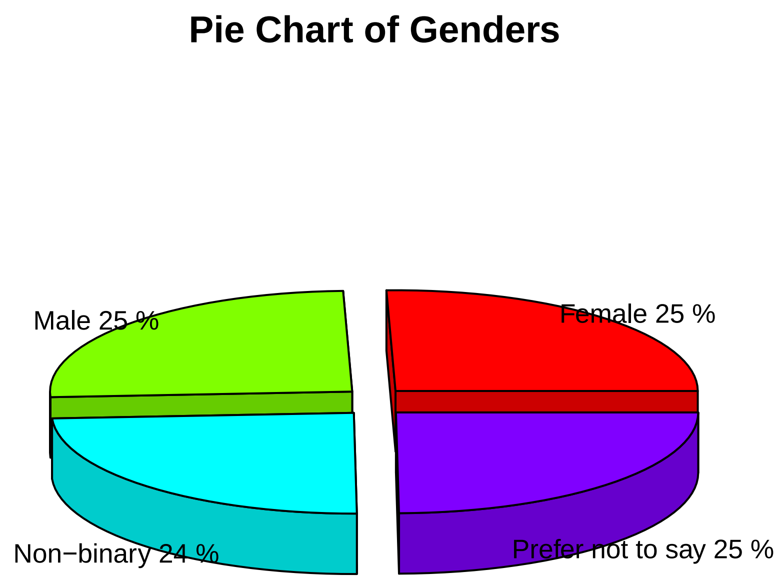
Gender Variable

Next, I want to focus on the gender variable showing the percentages of each gender in the dataset. This helps us understand a background of participants in the dataset.

```
gender_counts <- table(df$Gender)
gender_counts
```

```
       Female              Male        Non-binary Prefer not to say 
         1274              1270              1214              1242 
```

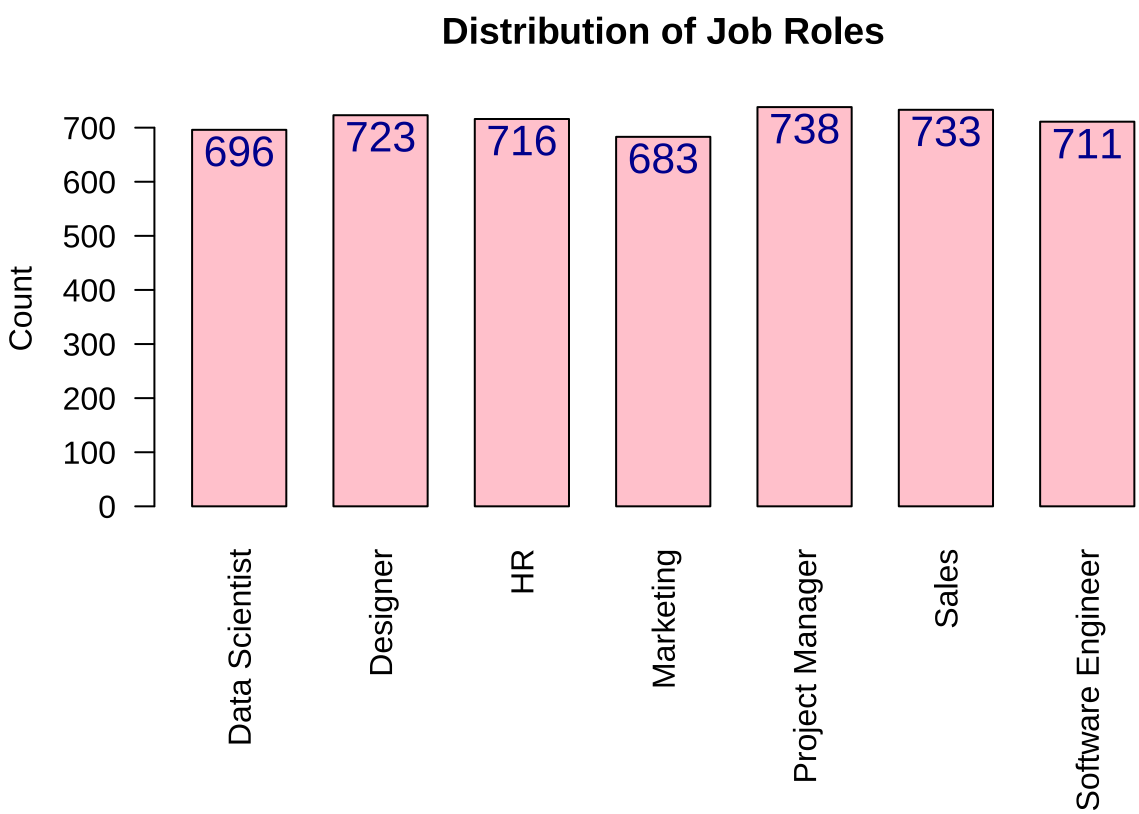
```
counts <- as.vector(gender_counts)

lbls <- names(gender_counts)
pct <- round(counts/sum(counts) * 100)
lbls <- paste(lbls, pct, "%", sep = " ")

suppressWarnings ({
pie3D(counts, labels = lbls, col = rainbow(length(lbls)),
      explode = 0.1,
      main = "Pie Chart of Genders",
      labelcex = 0.8,     # Adjust font size of labels
      labelradius = 1.2,  # Adjust distance of labels from center
      pos = 0)
})
```



Job Role

The boxplot of job roles reveals the distribution of employees across various positions

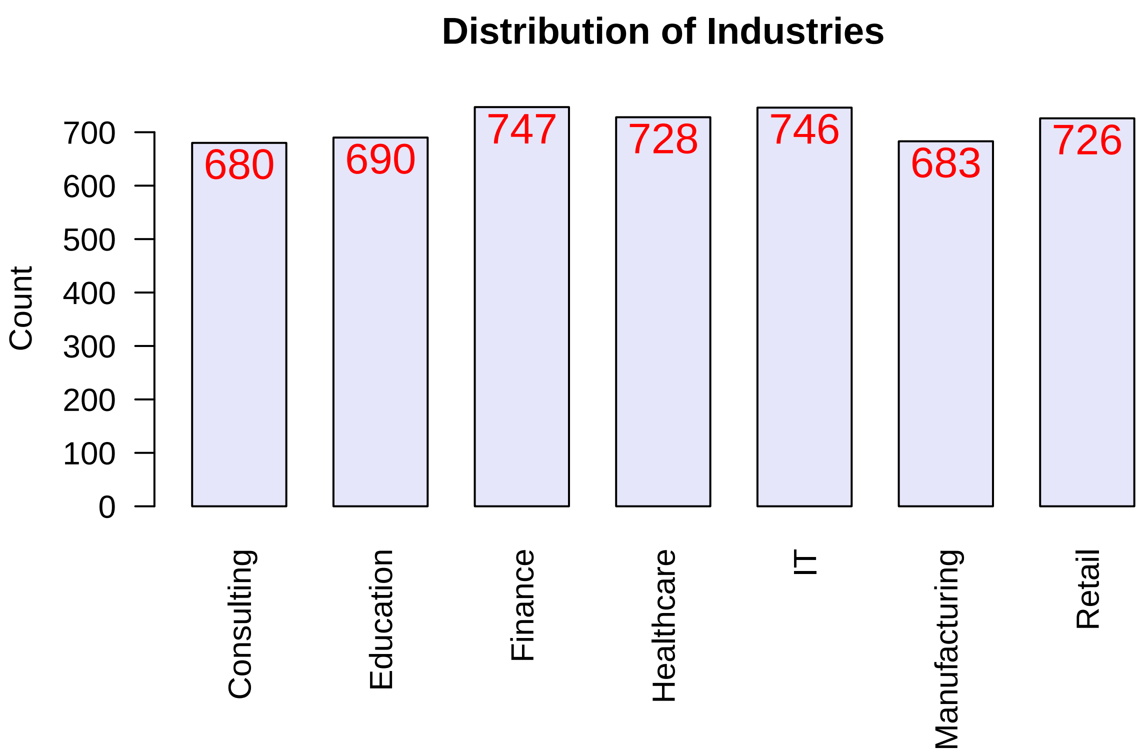
```
par(mar = c(8, 4, 4, 2))

job_roles <- table(df$Job_Role)

xx <- barplot(job_roles,
        main = "Distribution of Job Roles",
        ylab = "Count",
        col = "pink",
        las = 2,
        width = 0.5,   # Adjust bar width (default is 1)
        space = 0.5)

text(x = xx, y = job_roles - 40,
     label = as.character(job_roles),
    cex = 1.3,
    col = "darkblue"
    )
```



Industry

The boxplot of different industries reveals the distribution of industries with remote work.

```r
par(mar = c(8, 4, 4, 2))

industries <- table(df$Industry)

xx <- barplot(industries,
        main = "Distribution of Industries",
        ylab = "Count",
        col = "#E6E6FA",
        las = 2,
        width = 0.5,  # Adjust bar width (default is 1)
        space = 0.5)

text(x = xx, y = industries - 40,
     label = as.character(industries),
    cex = 1.3,
    col = "red"
    )
```



Mean Age, by Job Roles and Industry

```r
grouped_data <- group_by(df, Industry, Job_Role)
mean_age_by_industry <- summarise(grouped_data, Mean_Age = mean(Age, na.rm = TRUE, .groups = "drop"))

# Visual graph is only available to see in HTML/iOSlides presentation
#because of the interactive Plotly package.

plot_ly(data = mean_age_by_industry,
        x = ~Job_Role,
        y = ~Industry,
        z = ~Mean_Age,
        type = "heatmap",
        colorscale = "Viridis",
        colorbar = list(title = "Mean Age"),
        text = ~Mean_Age,  # Add text for annotations
        texttemplate = "%{text}",  # Display the text (mean ages) in the boxes
        hoverinfo = "text")  %>%
  layout(title = "Heatmap of Mean Age by Job Role and Industry",
         xaxis = list(title = "Job Role"),
         yaxis = list(title = "Industry"))
```

We can observe almost all job roles industries are dominated by employees in the age range of 39 - 43. Software Engineers and Designer mean age are all over 40 for each industry. Meanwhile HR has the most industries including the mean age under 40.

Work Location, Region

```
kable(table(df$Work_Location), caption = "Work Location")
```

Table 1: Work Location

Var1	Freq
Hybrid	1649
Onsite	1637
Remote	1714

```
kable(table(df$Region), caption = "Region")
```

Table 2: Region

Var1	Freq
Africa	860
Asia	829
Europe	840
North America	777
Oceania	867
South America	827

Stress Level, Mental Condition, Sleep Quality

```
kable(table(df$Stress_Level), caption = "Stress Level")
```

Table 3: Stress Level

Var1	Freq
High	1686
Low	1645
Medium	1669

```
kable(table(df$Mental_Health_Condition), caption = "Mental Health Condition")
```

Table 4: Mental Health Condition

Var1	Freq
Anxiety	1278
Burnout	1280
Depression	1246
None	1196

```
kable(table(df$Sleep_Quality), caption = "Sleep Quality")
```

Table 5: Sleep Quality

Var1	Freq
Average	1628
Good	1687
Poor	1685

Balanced Representation in the Dataset

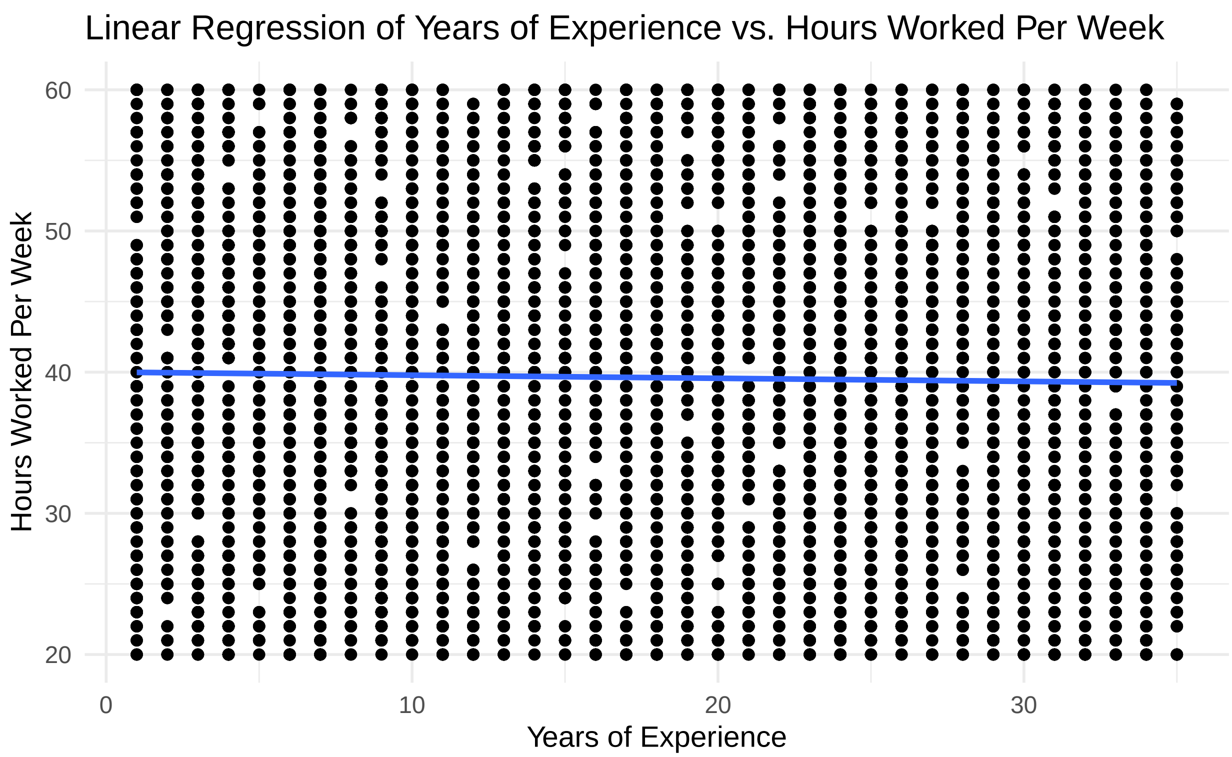
From diving deeper into a few of our variables in the dataset, we can see from the graphs and tables, each attribute in many of our variables is close to equal. With this diversed but equal representation, our analysis can more accurately reflect the experiences of each group. This balance allows us to explore deeper insights without bias, ensuring that all employee experiences are represented. Now, let's dive into visualizing relationships within the data and explore predictive models for further insights.

Objective and Problem Definition

This project explores key trends within the "Impact of Remote Work on Mental Health" dataset to identify factors influencing employee mental health. Specifically, I will investigate relationships between variables such as job role, stress level, and work-life balance to determine which factors most significantly impact mental well-being among on-site, hybrid and remote employees.

Years of Exp vs Hours Worked Per Week

```
ggplot(df, aes(Years_of_Experience, Hours_Worked_Per_Week)) +
  geom_point() +
  geom_smooth(method = "lm", se = F) +
  labs(title = "Linear Regression of Years of Experience vs. Hours Worked Per Week",
       x = "Years of Experience",
       y = "Hours Worked Per Week") +
  theme_minimal()
```



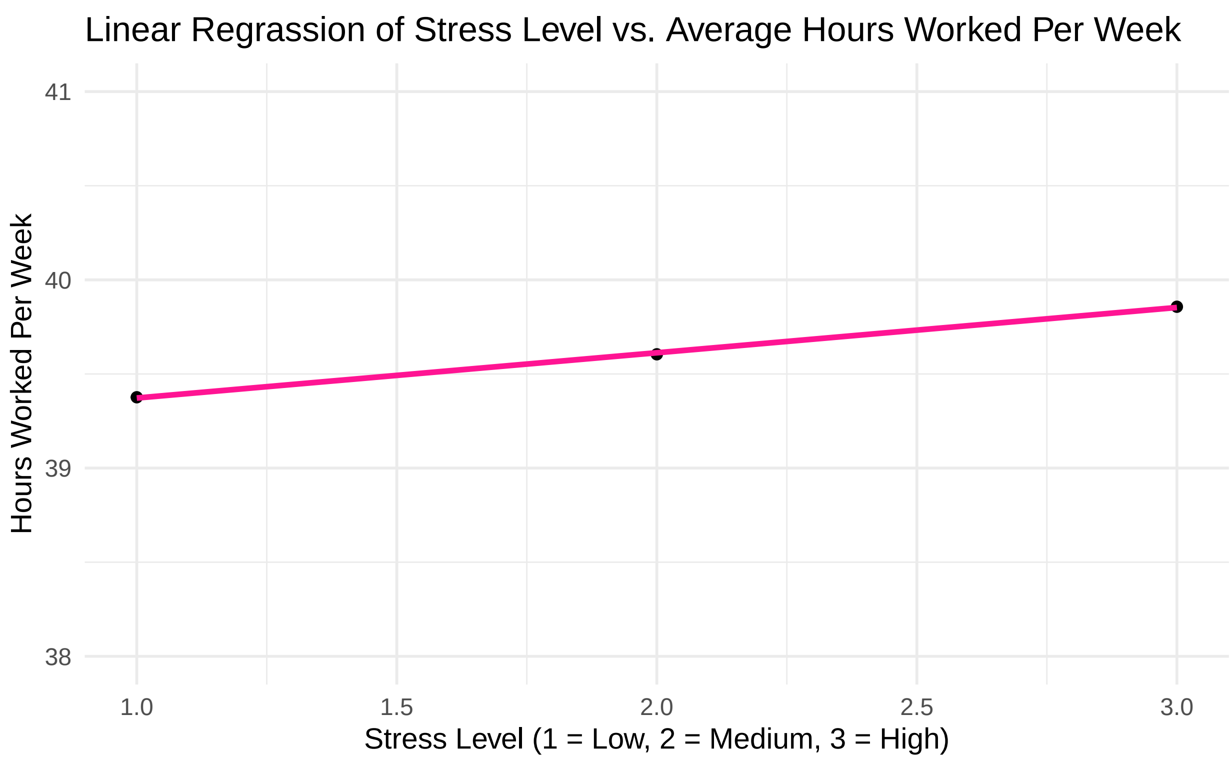

```
correlation <- cor(df$Years_of_Experience, df$Hours_Worked_Per_Week, use = "complete.obs")
print(correlation)
```

```
## [1] -0.01853681
```

Our calculated correlation coefficient is -0.0185 meaning there is minimal to no correlation between years of experience and hours worked, whether an employee has just a few years of experience or many years, it does not significantly affect the number of hours they work each week.

Stress Level vs Hours Worked

```
df <- df %>%
  mutate(Stress_Level_Num = case_when(
    Stress_Level == "Low" ~ 1,
    Stress_Level == "Medium" ~ 2,
    Stress_Level == "High" ~ 3,
  ))

average_hours <- df %>%
  group_by(Stress_Level_Num) %>%
  summarize(Avg_Hours_Worked = mean(Hours_Worked_Per_Week, na.rm = TRUE))

ggplot(average_hours, aes(x = Stress_Level_Num, y = Avg_Hours_Worked,)) +
  geom_point() +  # Adds the points
  geom_smooth(method = "lm", se = FALSE, color = "deeppink") +
   ylim(38,41) +
  labs(title = "Linear Regrassion of Stress Level vs. Average Hours Worked Per Week",
       x = "Stress Level (1 = Low, 2 = Medium, 3 = High)",
       y = "Hours Worked Per Week") +
  theme_minimal()
```



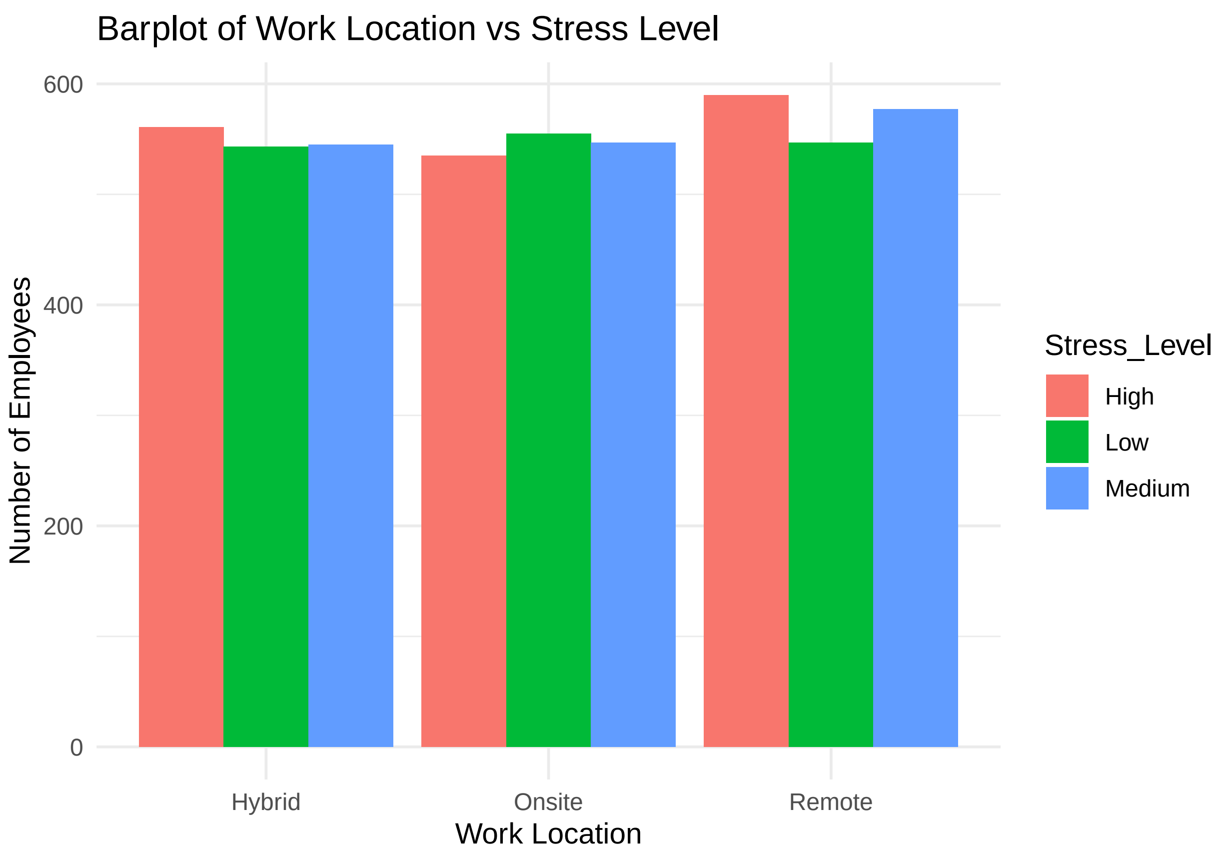

```
correlation <- cor(average_hours$Stress_Level_Num, average_hours$Avg_Hours_Worked)
print(correlation)
```

```
## [1] 0.9995765
```

Our calculated correlation coefficient is 0.99, the high correlation suggests that the variables are closely related. Employees who report higher stress levels tend to work more hours on average.

Work Location vs Stress Level

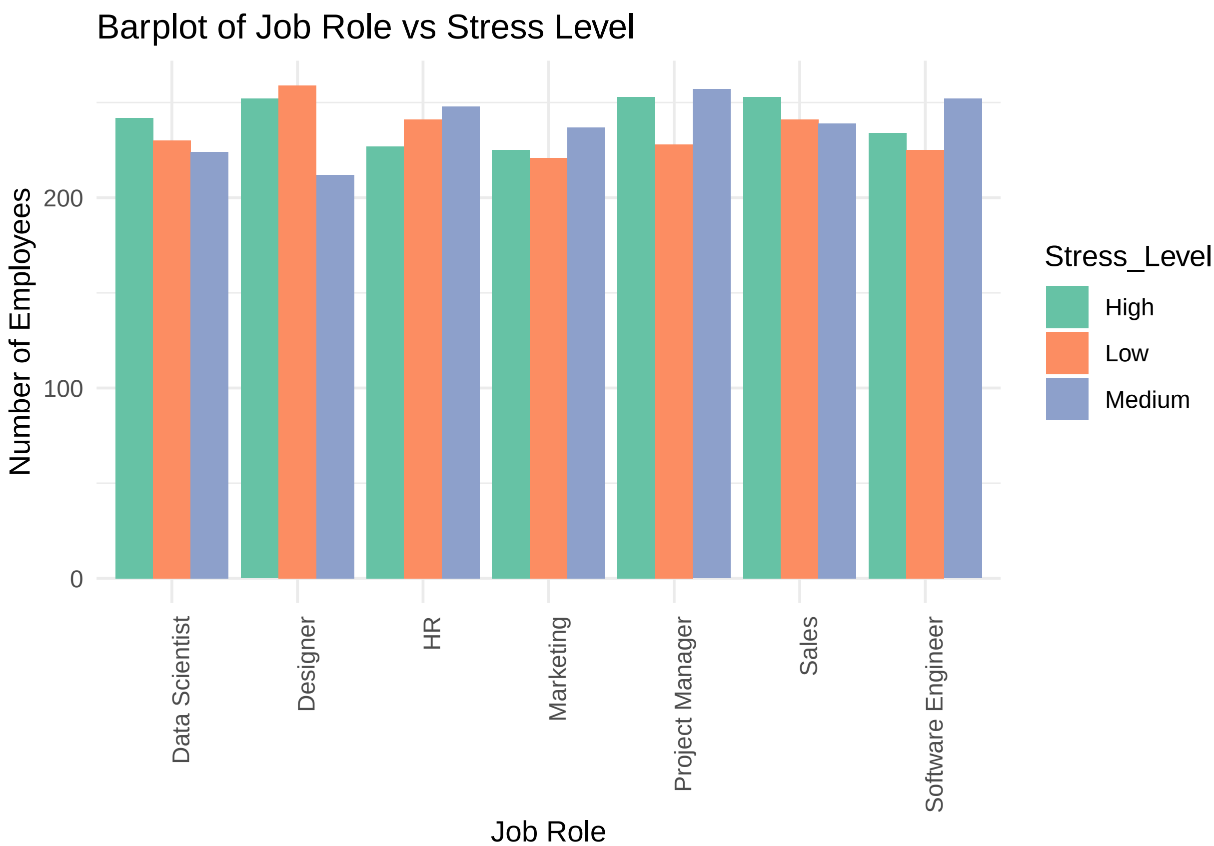
```
location_stress <- df %>%
  group_by(Work_Location, Stress_Level) %>%
  summarize(count = n())

location_stress %>%
  ggplot(aes(Work_Location, count, fill = Stress_Level)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Barplot of Work Location vs Stress Level",
       x = "Work Location",
       y = "Number of Employees") +
  theme_minimal()
```



We can observe employees working in a hybrid or remote model experience higher stress levels.

Job Role vs Stress Level

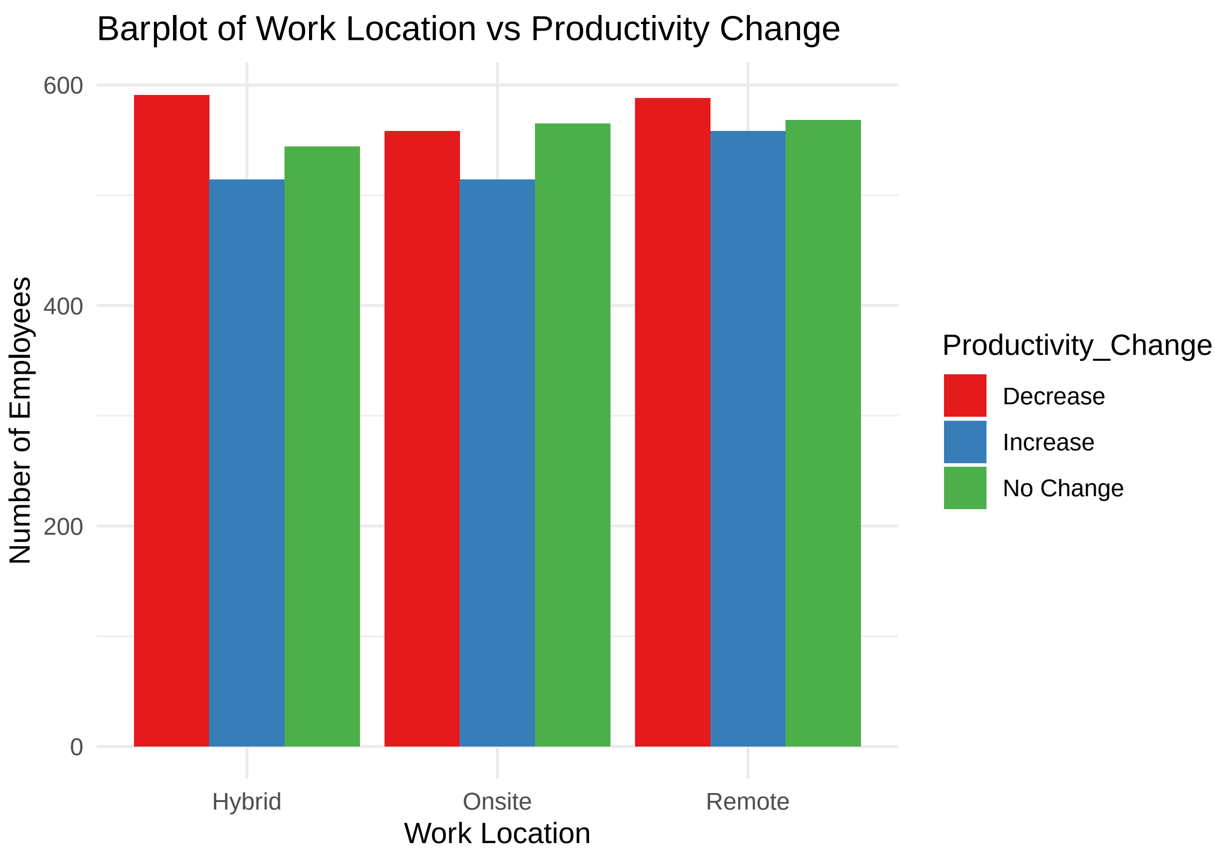
```
job_stress <- df %>%
  group_by(Job_Role, Stress_Level) %>%
  summarize(count = n())

job_stress %>%
  ggplot(aes(Job_Role, count, fill = Stress_Level)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Barplot of Job Role vs Stress Level",
       x = "Job Role",
       y = "Number of Employees") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set2") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Employees of jobs including data scientists and sales tend to have a majority stress level of high. Employees of jobs including HR, marketing, and software engineers have a majority stress level of medium while designers and project managers have a majority stress level of low.

Rate of Productivity vs Work Location

```
location_productivity <- df %>%
  group_by(Work_Location, Productivity_Change) %>%
  summarize(count = n())

location_productivity %>%
  ggplot(aes(Work_Location, count, fill = Productivity_Change)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Barplot of Work Location vs Productivity Change",
       x = "Work Location",
       y = "Number of Employees") +
  theme_minimal() +
  scale_fill_brewer(palette = "Set1")
```

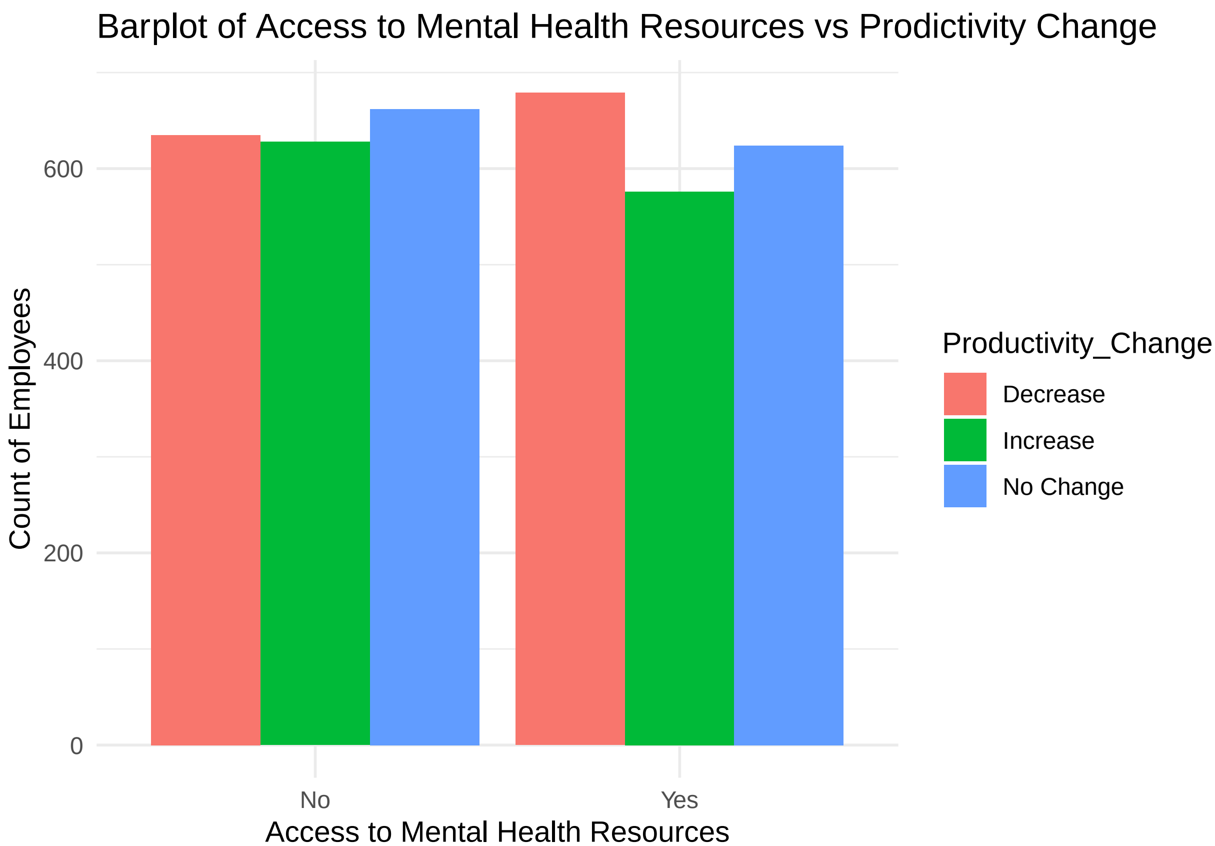


Employees working in a remote or hybrid work model tend to experience a decrease in their productivity.

Access to Mental Help vs Productivity

Firse, we filter out the employees who have no mental health conditions.

```r
## Filter out employees with no mental health condition
filtered_data <- df %>%
  filter(Mental_Health_Condition != "None")

ggplot(filtered_data, aes(x = Access_to_Mental_Health_Resources, fill = Productivity_Change)) +
  geom_bar(position = "dodge") +
  labs(title = "Barplot of Access to Mental Health Resources vs Prodictivity Change",
       x = "Access to Mental Health Resources",
       y = "Count of Employees") +
  theme_minimal()
```



Among employees who have access to mental health resources, the majority report a decrease in productivity. In contrast, most employees without access to mental health resources exhibit no change in productivity. This suggests that while mental health resources are crucial for support, they may not directly correlate with productivity gains in the short term. Alternatively, it could indicate that those already struggling with productivity may be more likely to seek out these resources.

Work Life Balance vs Job Role

On a scale from 1-5, we will calculate and show the mean work life balance rating for each job role.

```
group_data <- group_by(df, Job_Role, Work_Location)
mean_rate_by_job <- summarise(group_data, Mean_WorkBalanceRate = mean(Work_Life_Balance_Rating, na.rm =

# Visual graph is only available to see in HTML/iOSlides presentation
#because of the interactive Plotly package.

plot_ly(data = mean_rate_by_job,
        x = ~Work_Location,
        y = ~Job_Role,
        z = ~Mean_WorkBalanceRate,
        type = "heatmap",
        colorscale = "Viridis",
        colorbar = list(title = "Mean Work Life Balance Rating"),
        text = ~Mean_WorkBalanceRate,  # Add text for annotations
        texttemplate = "%{text}",  # Display the text (mean ages) in the boxes
        hoverinfo = "text")  %>%
  layout(title = "Heatmap of Mean Work Life Balance Rating by Work Location and Job Role",
         xaxis = list(title = "Work Location"),
         yaxis = list(title = "Job Role"))
```

Remote software engineers have the highest average work life balance at 3.2, while remote employees working in marketing have the lowest average work life balance rate at 2.7.

Physical Activity and Sleep Quality

This table shows the most common values for physical activity and sleep quality for employees in each work location.

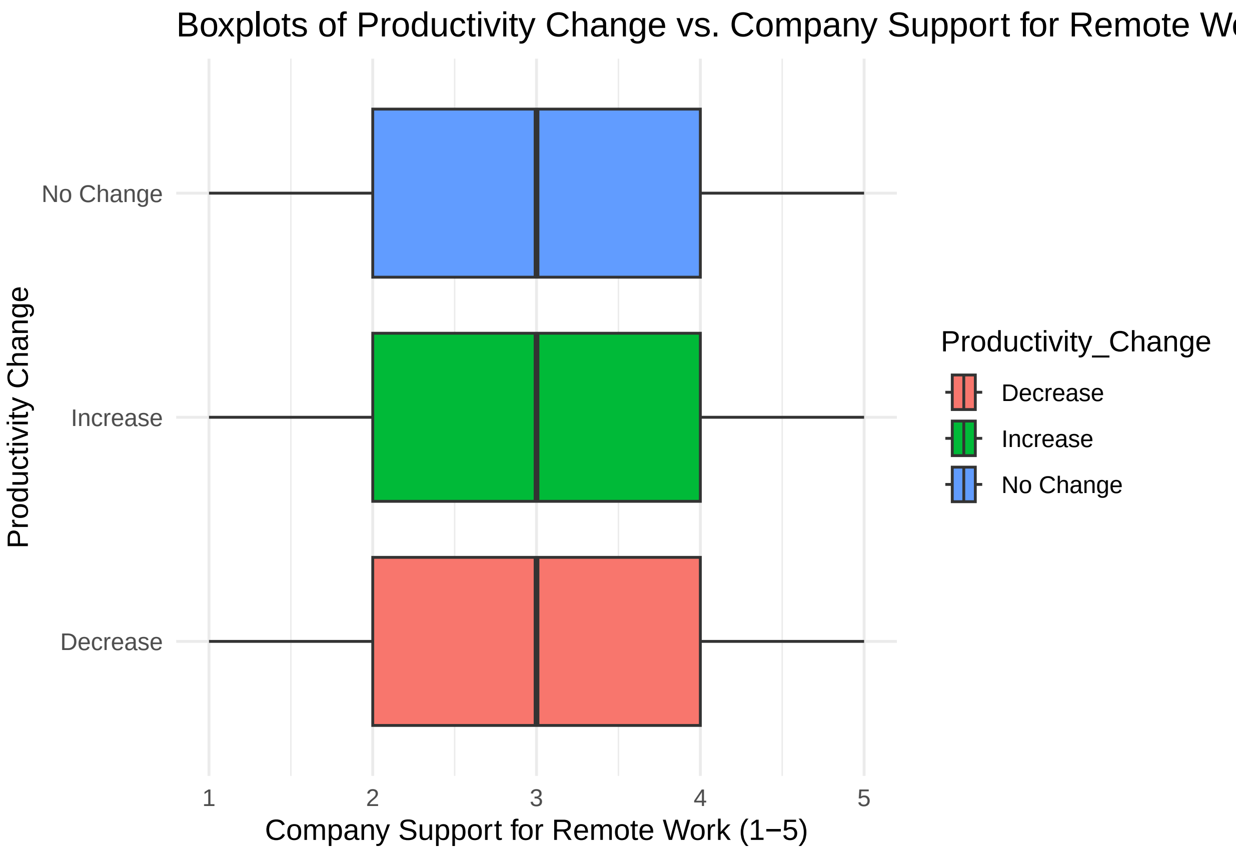
```r
# Define a function to get the mode
get_mode <- function(x) {
  unique_x <- unique(x)
  unique_x[which.max(tabulate(match(x, unique_x)))]
}

# Group by Work_Location and calculate mode for Physical_Activity and Sleep_Quality
health_summary <- df %>%
  group_by(Work_Location) %>%
  summarize(
    Physical_Activity_Mode = get_mode(Physical_Activity),
    Sleep_Quality_Mode = get_mode(Sleep_Quality)
  )

health_summary
```

```
## # A tibble: 3 x 3
##   Work_Location Physical_Activity_Mode Sleep_Quality_Mode
##   <chr>         <chr>                  <chr>
## 1 Hybrid        Weekly                 Good
## 2 Onsite        Weekly                 Poor
## 3 Remote        Daily                  Average
```

Hybrid employees work out weekly with good sleep quality. Onsite employees work out weekly but have poor sleep quality. Remote employees work out daily with average sleep quality.

Productivity vs Company Support

```
ggplot(df, aes(x = Company_Support_for_Remote_Work, y = Productivity_Change, fill = Productivity_Change
  geom_boxplot() +
  labs(title = "Boxplots of Productivity Change vs. Company Support for Remote Work",
    x = "Company Support for Remote Work (1-5)",
    y = "Productivity Change") +
  theme_minimal()
```



We can observe there is no correlation between productivity change and the given company support rating.

Age, Hours Worked & Satisfaction Level

```r
# Visual graph is only available to see in HTML/iOSlides presentation
#because of the interactive Plotly package.

plot_ly(data = df, x = ~Age, y = ~Hours_Worked_Per_Week, z = ~Satisfaction_with_Remote_Work,
        color = ~Job_Role, colors = "plasma", size = I(5)) %>%
  add_markers() %>%
  layout(scene = list(
    xaxis = list(title = 'Age'),
    yaxis = list(title = 'Hours Worked Per Week'),
    zaxis = list(title = 'Satisfaction Level'),
    title = "3D Scatter Plot: Age, Hours Worked, and Satisfaction Level by Job Role"
  ))
```

This 3D scatter plot illustrates the relationship between employee age, hours worked per week, and satisfaction level regarding remote work, with points color-coded by job role. The distribution of points highlights potential trends, such as how satisfaction levels vary across different age groups and workloads.

Conclusion

The analysis of the "Impact of Remote Work on Mental Health" dataset highlights several trends in employee demographics and well-being. Employees aged 39 to 43 dominate various job roles, particularly in Software Engineering and Design. The dataset reflects a balanced representation across job roles, enhancing the credibility of our insights.

A strong correlation (0.99) indicates that higher stress levels are associated with longer working hours, while years of experience show minimal correlation (-0.0185) with hours worked. Job roles like Data Scientists and Sales exhibit higher stress levels, whereas remote and hybrid employees tend to report decreased productivity.

Interestingly, access to mental health resources does not guarantee productivity gains for those struggling, and remote Software Engineers report the highest work-life balance. Overall, these findings reveal the complex relationships between job roles, stress, productivity, and mental health resources, paving the way for further exploration of these dynamics.