

Data 607 Week 1

Benjamin Bravo

2026-02-01

Overview:

This dataset examines daily work-from-home behaviors and how they relate to employee burnout and productivity. It includes about 1,800 daily records with information on work hours, screen time, meetings, breaks, sleep, and burnout levels across both weekdays and weekends. The data can be used to study burnout risk and how work habits affect productivity and well-being.

Link: <https://www.kaggle.com/datasets/sonalshinde123/work-from-home-employee-burnout-dataset?resource=download>

```
install.packages("tidyverse")

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

library(tidyverse)

## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.6
## vforcats   1.0.1     v stringr   1.6.0
## v ggplot2   4.0.1     v tibble    3.3.0
## v lubridate 1.9.4     v tidyr    1.3.2
## v purrr    1.2.0

## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()   masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

data_url <- "https://raw.githubusercontent.com/bb2955/Data-607/main/work_from_home_burnout_dataset.csv"

df <- readr::read_csv(data_url, show_col_types = FALSE)

install.packages("conflicted")

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

library(conflicted)

conflict_prefer_all("dplyr", quiet = TRUE)

problems(df)

## # A tibble: 0 x 5
## # i 5 variables: row <int>, col <int>, expected <chr>, actual <chr>, file <chr>
```

```

colnames(df)

## [1] "user_id"           "day_type"          "work_hours"
## [4] "screen_time_hours" "meetings_count"   "breaks_taken"
## [7] "after_hours_work"  "sleep_hours"       "task_completion_rate"
## [10] "burnout_score"     "burnout_risk"

head(df)

## # A tibble: 6 x 11
##   user_id day_type work_hours screen_time_hours meetings_count breaks_taken
##   <dbl> <chr>      <dbl>        <dbl>            <dbl>            <dbl>
## 1 1     Weekday    9.59        11.9             4                2
## 2 1     Weekend    7.38        10.3             4                1
## 3 1     Weekend    6.31        8.92            1                2
## 4 1     Weekday    8.34        10.7             4                1
## 5 1     Weekend    6.97        9.83            1                2
## 6 1     Weekday    7.24        9.09            1                4
## # i 5 more variables: after_hours_work <dbl>, sleep_hours <dbl>,
## #   task_completion_rate <dbl>, burnout_score <dbl>, burnout_risk <chr>
glimpse(df)

## Rows: 1,800
## Columns: 11
## $ user_id           <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2~
## $ day_type          <chr> "Weekday", "Weekend", "Weekend", "Weekday", "Week~
## $ work_hours         <dbl> 9.59, 7.38, 6.31, 8.34, 6.97, 7.24, 8.09, 7.15, 8~
## $ screen_time_hours <dbl> 11.86, 10.33, 8.92, 10.70, 9.83, 9.09, 11.64, 9.9~
## $ meetings_count     <dbl> 4, 4, 1, 4, 1, 1, 6, 3, 1, 0, 3, 3, 0, 3, 1, 0, 3~
## $ breaks_taken       <dbl> 2, 1, 2, 1, 2, 4, 3, 4, 2, 4, 5, 3, 5, 1, 5, 3, 1~
## $ after_hours_work   <dbl> 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0~
## $ sleep_hours        <dbl> 7.55, 6.69, 8.87, 8.13, 5.85, 7.53, 5.04, 5.89, 6~
## $ task_completion_rate <dbl> 91.2, 82.0, 80.6, 70.0, 67.1, 69.1, 58.4, 46.9, 4~
## $ burnout_score      <dbl> 19.17, 29.70, 32.93, 45.47, 51.61, 54.16, 68.83, ~
## $ burnout_risk        <chr> "Low", "Low", "Low", "Low", "Low", "Low", ~

summary(df)

## #> #> user_id      day_type      work_hours    screen_time_hours
## #> #> Min.   : 1.00  Length:1800    Min.   : 3.000  Min.   : 4.510
## #> #> 1st Qu.: 45.75 Class :character  1st Qu.: 4.430  1st Qu.: 7.240
## #> #> Median : 90.50 Mode  :character  Median : 6.445  Median : 9.210
## #> #> Mean   : 90.50                  Mean   : 6.515  Mean   : 9.271
## #> #> 3rd Qu.:135.25                  3rd Qu.: 8.510  3rd Qu.:11.310
## #> #> Max.   :180.00                  Max.   :12.170  Max.   :15.700
## #> #> meetings_count  breaks_taken  after_hours_work sleep_hours
## #> #> Min.   : 0.000  Min.   :1.000  Min.   :0.0000  Min.   : 4.500
## #> #> 1st Qu.: 1.000  1st Qu.:2.000  1st Qu.:0.0000  1st Qu.: 6.280
## #> #> Median : 2.000  Median :3.000  Median :0.0000  Median : 6.990
## #> #> Mean   : 1.941  Mean   :3.029  Mean   :0.3589  Mean   : 6.996
## #> #> 3rd Qu.: 3.000  3rd Qu.:4.000  3rd Qu.:1.0000  3rd Qu.: 7.750
## #> #> Max.   :10.000  Max.   :5.000  Max.   :1.0000  Max.   :10.800
## #> #> task_completion_rate burnout_score  burnout_risk
## #> #> Min.   : 40.00    Min.   :  2.50  Length:1800
## #> #> 1st Qu.: 62.30    1st Qu.: 25.37  Class :character

```

```

## Median : 74.50      Median : 39.27    Mode  :character
## Mean   : 72.31      Mean   : 44.01
## 3rd Qu.: 83.70      3rd Qu.: 58.20
## Max.   :107.20      Max.   :143.92

df_clean <- df %>%
  rename(
    id = user_id,
    num_meetings = meetings_count,
    num_breaks = breaks_taken,
    burnout_score = burnout_score,    # keep same but shown for clarity
    burnout_risk = burnout_risk
  )

df_clean <- df_clean %>%
  mutate(
    day_type = case_when(
      day_type %in% c("W", "wkday", "weekday") ~ "Weekday",
      day_type %in% c("E", "wknd", "weekend") ~ "Weekend",
      TRUE ~ as.character(day_type)
    ),
    after_hours_work = case_when(
      after_hours_work %in% c("Y", "Yes", "1") ~ "Yes",
      after_hours_work %in% c("N", "No", "0") ~ "No",
      TRUE ~ as.character(after_hours_work)
    ),
    burnout_risk = case_when(
      burnout_risk %in% c("L", "low") ~ "Low",
      burnout_risk %in% c("M", "med", "medium") ~ "Medium",
      burnout_risk %in% c("H", "high") ~ "High",
      TRUE ~ as.character(burnout_risk)
    ),
    day_type = factor(day_type),
    after_hours_work = factor(after_hours_work),
    burnout_risk = factor(burnout_risk, levels = c("Low", "Medium", "High"))
  )

df %>% summarise(
  day_type_vals = paste(sort(unique(day_type)), collapse = ", "),
  after_hours_vals = paste(sort(unique(after_hours_work)), collapse = ", "),
  burnout_risk_vals = paste(sort(unique(burnout_risk)), collapse = ", ")
)

## # A tibble: 1 x 3
##   day_type_vals     after_hours_vals burnout_risk_vals
##   <chr>            <chr>           <chr>
## 1 Weekday, Weekend 0, 1             High, Low, Medium
table(df_clean$day_type)

##
## Weekday Weekend
##     876      924
table(df_clean$after_hours_work)

##

```

```

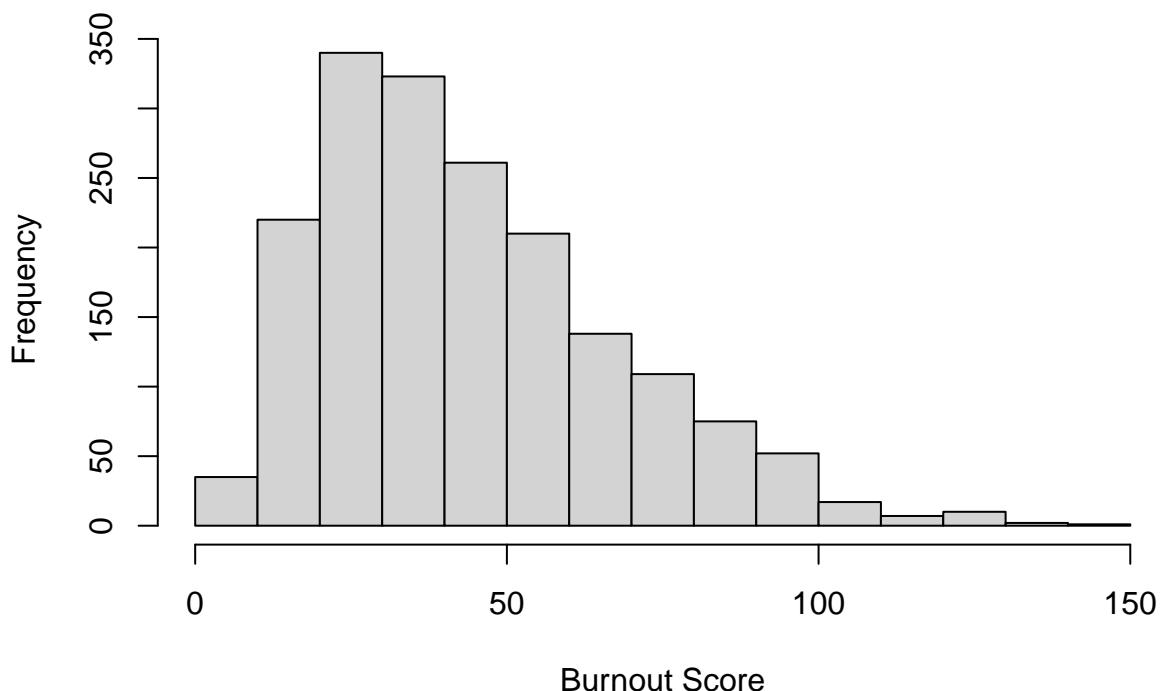
##   No  Yes
## 1154 646





```

Distribution of Burnout Score

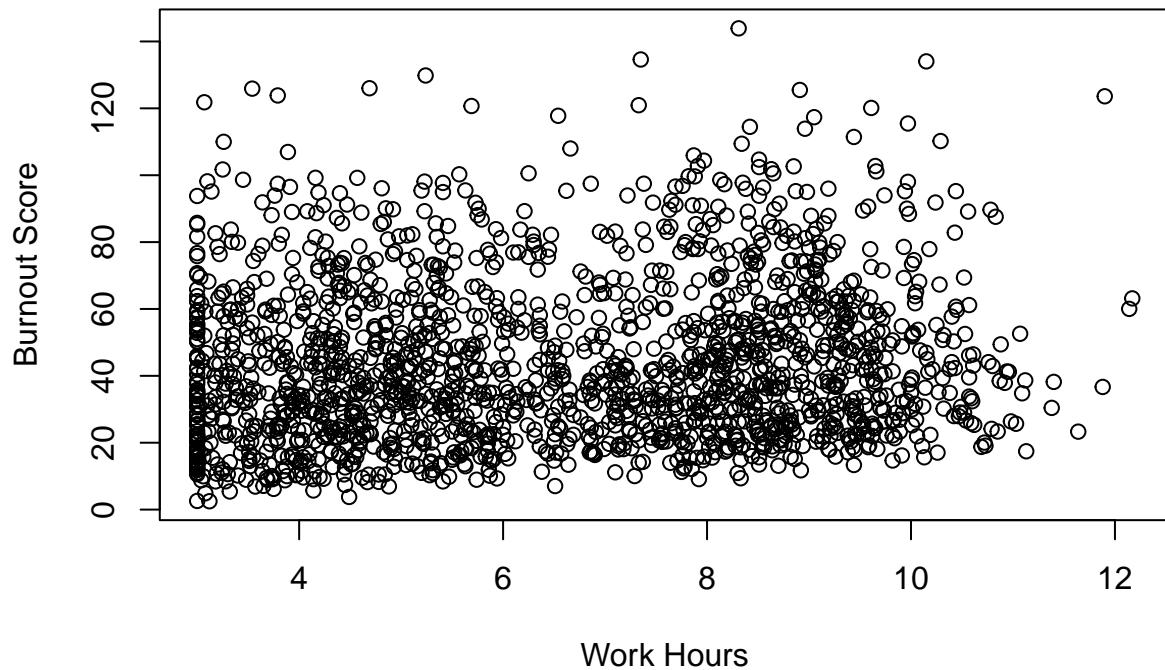


```

plot(
  df_clean$work_hours,
  df_clean$burnout_score,
  xlab = "Work Hours",
  ylab = "Burnout Score",
  main = "Work Hours vs Burnout Score"
)

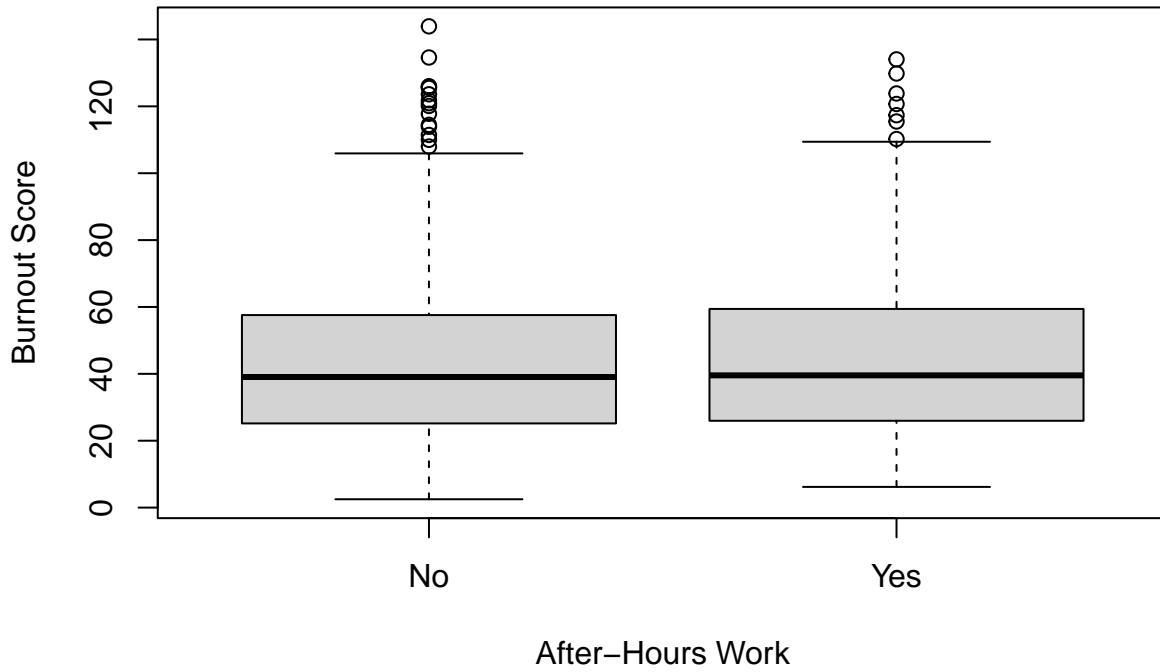
```

Work Hours vs Burnout Score



```
boxplot(  
  burnout_score ~ after_hours_work,  
  data = df_clean,  
  xlab = "After-Hours Work",  
  ylab = "Burnout Score",  
  main = "Burnout Score by After-Hours Work"  
)
```

Burnout Score by After-Hours Work



Conclusion:

The exploratory analysis suggests that working after hours and longer work hours are associated with slightly higher burnout scores, but these relationships show substantial variability. This indicates that burnout is influenced by multiple factors, and future work should include additional variables and more advanced models to better understand and predict burnout risk. To update and extend this work, additional variables such as job role, workload intensity (task difficulty or deadlines), and stress or mental health indicators would be useful because burnout is often influenced by more than just hours worked. Including manager support, flexibility level, and long-term trends over time would also help explain the variability seen in the graphs and lead to more accurate burnout prediction and interpretation.