Newdata

Chang Lu

2025-03-27

```
library(readr)
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(lubridate)
```

```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(stringr)
library(purrr)
library(readxl)
```

```
# Computing Total_Quit_Time_Computed
submit_time_cols <- names(combined_df)[str_detect(names(combined_df), "-T_Page Submit")]

combined_df$Total_Quit_Time_Computed <- combined_df %>%
  select(all_of(submit_time_cols)) %>%
  mutate_all(as.numeric) %>%
  rowSums(na.rm = TRUE)


combined_df %>% select(...1, Total_Quit_Time_Computed) %>% head()
```

```
## # A tibble: 6 x 2
##   ...1      Total_Quit_Time_Computed
##   <chr>                        <dbl>
## 1 Student 1                     38.9
## 2 Student 2                   1243.
## 3 Student 3                   1209.
## 4 Student 4                    172.
## 5 Student 5                   1495.
## 6 Student 6                   2465.
```

```r
length(combined_df$Total_Quit_Time_Computed)
```

```
## [1] 32
```

```r
# Rename demographic groups
combined_df <- combined_df %>%
  rename(
    Accommodations       = Q21,
    Gender               = Q22,
    Ethnicity            = Q23,
    English_Proficiency  = Q24,
    Country_You          = Q25_1,
    Country_Mother       = Q25_2,
    Country_Father       = Q25_3,
    Home_Language        = Q69
  ) %>%
  select(-Q68)  # Drop Q68 for a few people answered
```

Normality Check

```r


# # Accomodations
# group_yes <- combined_df %>% filter(Accommodations == "Yes") %>% pull(Total_Quit_Time_Computed)
# group_no  <- combined_df %>% filter(Accommodations == "No") %>% pull(Total_Quit_Time_Computed)
#
# # Shapiro-Wilk test
# shapiro.test(group_yes)
# shapiro.test(group_no)
```

```r


# par(mfrow = c(2, 2))  # Plot layout
#
# # Histogram
# hist(group_yes, main = "Accommodations: YES", xlab = "Total Time", col = "skyblue")
# hist(group_no, main = "Accommodations: NO", xlab = "Total Time", col = "salmon")
#
# # Q-Q plots
# qqnorm(group_yes); qqline(group_yes, col = "blue")
# qqnorm(group_no); qqline(group_no, col = "red")
```

Accomodations Error in shapiro.test(group_yes) : sample size must be between 3 and 5000 -> Mann–Whitney U test

```r
# Group sizes
table(combined_df$Gender)
```

Gender

```
##
## Female   Male
##     23      8
```

```r
# Normality check
by(combined_df$Total_Quit_Time_Computed, combined_df$Gender, shapiro.test)
```

```
## combined_df$Gender: Female
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.75886, p-value = 9.016e-05
##
## ------------------------------------------------------------
## combined_df$Gender: Male
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.95465, p-value = 0.7579
```

We tested whether total quit time differed by gender. The Female group violated the assumption of normality (Shapiro-Wilk $p < 0.001$), so we used a non-parametric Mann–Whitney U test.

```r
table(combined_df$Ethnicity)
```

Race

```
##
##                      Asian Black or African American        Hispanic or Latino
##                         10                         1                         5
##                      Other                     White
##                          2                        13
```

```r
# Kruskal-Wallis test (non-parametric ANOVA)
kruskal.test(Total_Quit_Time_Computed ~ Ethnicity, data = combined_df)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Total_Quit_Time_Computed by Ethnicity
## Kruskal-Wallis chi-squared = 1.0582, df = 4, p-value = 0.9008
```

A Kruskal-Wallis test revealed no significant difference in total quit time across ethnic groups ($\chi^2(4) = 1.06$, p = 0.901).

```r
# Clean and recode English proficiency
combined_df <- combined_df %>%
  mutate(English_Proficiency = case_when(
    str_trim(English_Proficiency) %in% c("Native", "Native Speaker") ~ "Native",
    str_trim(English_Proficiency) == "Fluent" ~ "Fluent",
    TRUE ~ English_Proficiency  # keep as-is in case there's other unexpected values
  ))

table(combined_df$English_Proficiency)
```

English Proficiency

```
##
## Fluent Native
##     13     18
```

```r
wilcox.test(Total_Quit_Time_Computed ~ English_Proficiency, data = combined_df)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  Total_Quit_Time_Computed by English_Proficiency
## W = 113, p-value = 0.8902
## alternative hypothesis: true location shift is not equal to 0
```

A Wilcoxon rank-sum test found no significant difference in total quit time between Fluent and Native English speakers (W = 113, p = 0.89)

```r
# You
combined_df <- combined_df %>%
  mutate(Born_in_US = ifelse(Country_You == "United States", "Yes", "No"))

wilcox.test(Total_Quit_Time_Computed ~ Born_in_US, data = combined_df)
```

Born in US

```
##
##  Wilcoxon rank sum exact test
##
## data:  Total_Quit_Time_Computed by Born_in_US
## W = 92, p-value = 0.729
## alternative hypothesis: true location shift is not equal to 0
```

A Wilcoxon rank-sum test found no significant difference in total quit time between students born in the U.S. and those born elsewhere (W = 92, p = 0.73).

```
# Mother
combined_df <- combined_df %>%
  mutate(Born_in_US_M = ifelse(Country_Mother == "United States", "Yes", "No"))

wilcox.test(Total_Quit_Time_Computed ~ Born_in_US_M, data = combined_df)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  Total_Quit_Time_Computed by Born_in_US_M
## W = 156, p-value = 0.09266
## alternative hypothesis: true location shift is not equal to 0
```

A Wilcoxon rank-sum test comparing total quit time by mother's country of birth showed a trend toward significance (W = 156, p = 0.093), but did not reach conventional significance levels. It may need larger sample size to support.

```
# Father
combined_df <- combined_df %>%
  mutate(Born_in_US_F = ifelse(Country_Father == "United States", "Yes", "No"))

wilcox.test(Total_Quit_Time_Computed ~ Born_in_US_F, data = combined_df)
```

```
##
##  Wilcoxon rank sum exact test
##
## data:  Total_Quit_Time_Computed by Born_in_US_F
## W = 149, p-value = 0.2107
## alternative hypothesis: true location shift is not equal to 0
```

A Wilcoxon rank-sum test showed no significant difference in total quit time based on father's country of birth (W = 149, p = 0.211).

```
combined_df <- combined_df %>%
  mutate(English_Home = ifelse(Home_Language == "English", "Yes", "No"))

wilcox.test(Total_Quit_Time_Computed ~ English_Home, data = combined_df)
```

Home Language

```
##
##  Wilcoxon rank sum exact test
##
## data:  Total_Quit_Time_Computed by English_Home
## W = 152, p-value = 0.1297
## alternative hypothesis: true location shift is not equal to 0
```

A Wilcoxon rank-sum test showed no significant difference in total quit time between students who spoke English at home and those who did not (W = 152, p = 0.130).

To qiuyi

Hi Qiuyi, when comparing total quit time across demographic groups, here's how to decide between using a t-test or the Mann–Whitney U test:

Use t-test if: Both groups have at least 10 observations and the distribution of total quit time is approximately normal in both groups (check via Shapiro-Wilk or Q-Q plot)(I didnt draw plots) \

Use Mann–Whitney U test if: 1.One or both groups violate normality

2.Sample size is small (especially < 10)