25DataAnalysis

Chang Lu

2025-04-30

```
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(ggplot2)
library(readxl)
library(rstatix)
```

```
## Warning: package 'rstatix' was built under R version 4.4.3

##
## Attaching package: 'rstatix'

## The following object is masked from 'package:stats':
##
##     filter
```

```
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.4.3
```

```
# Read the data

combined_df <- read_excel("New - MAMS and Dental Combined.xlsx")
```

```
## New names:
## * `` -> `...1`
## * `E.` -> `E....10`
## * `E.` -> `E....12`
```

```r
head(combined_df)
```

```
## # A tibble: 6 x 29
##   ...1      Central diabetes ins~1 Decreased conduction~2 Mitral valve stenosi~3
##   <chr>     <chr>                  <chr>                  <chr>
## 1 Student   Q1                     Q2                     Q3
## 2 Student 1 1.0                    0.0                    0.0
## 3 Student 2 1.0                    1.0                    0.0
## 4 Student 3 0.0                    1.0                    0.0
## 5 Student 4 0.0                    0.0                    0.0
## 6 Student 5 1.0                    0.0                    1.0
## # i abbreviated names: 1: ‘Central diabetes insipidus‘,
## #   2: ‘Decreased conduction rate along the bundle branches‘,
## #   3: ‘Mitral valve stenosis‘
## # i 25 more variables:
## #   ‘Decreased pulmonary capillary hydrostatic fluid pressure‘ <chr>,
## #   Oxytocin <chr>, ‘Graves’ disease‘ <chr>, B. <chr>,
## #   ‘Increased serum aldosterone concentration‘ <chr>, E....10 <chr>, ...
```

```r
colnames(combined_df)
```

```
##  [1] "...1"
##  [2] "Central diabetes insipidus"
##  [3] "Decreased conduction rate along the bundle branches"
##  [4] "Mitral valve stenosis"
##  [5] "Decreased pulmonary capillary hydrostatic fluid pressure"
##  [6] "Oxytocin"
##  [7] "Graves’ disease"
##  [8] "B."
##  [9] "Increased serum aldosterone concentration"
## [10] "E....10"
## [11] "Arterial O2 concentration"
## [12] "E....12"
## [13] "Blocked urethra"
## [14] "Excess maternal androgens"
## [15] "The elastic recoil of the stretched arterial walls provides the force to continue blood flow in
## [16] "Mutations that result in inactive IGF-1 receptors"
## [17] "A decrease in Ca2+ resorption from bone"
## [18] "Absence of a Y chromosome"
## [19] "Testosterone stimulates GnRH from the hypothalamus"
## [20] "Plasma angiotensin II concentration increases"
## [21] "Its production is enhanced by cortisol."
## [22] "Total Score"
## [23] "Accomodations"
## [24] "Sex"
## [25] "Race/Ethnicity"
## [26] "English Proficiency"
## [27] "Born USA"
## [28] "Home Language"
## [29] "Age arrive USA"
```

```r
combined_df_new <- combined_df %>%
  slice(-1) %>%
  rename(student = 1) %>%
  mutate(across(c(`Total Score`, Accomodations, Sex, `Race/Ethnicity`,
                `English Proficiency`, `Born USA`, `Home Language`,
                `Age arrive USA`),
               ~ parse_number(as.character(.))))%>%
  rename(Total_Score = `Total Score`,
         Race_ethnicity  = `Race/Ethnicity`,
         English_prof    = `English Proficiency`,
         Born_usa        = `Born USA`,
         Home_language   = `Home Language`)

head(combined_df_new)
```

```
## # A tibble: 6 x 29
##   student   Central diabetes ins~1 Decreased conduction~2 Mitral valve stenosi~3
##   <chr>     <chr>                  <chr>                  <chr>
## 1 Student 1 1.0                    0.0                    0.0
## 2 Student 2 1.0                    1.0                    0.0
## 3 Student 3 0.0                    1.0                    0.0
## 4 Student 4 0.0                    0.0                    0.0
## 5 Student 5 1.0                    0.0                    1.0
## 6 Student 6 1.0                    1.0                    0.0
## # i abbreviated names: 1: ‘Central diabetes insipidus‘,
## #   2: ‘Decreased conduction rate along the bundle branches‘,
## #   3: ‘Mitral valve stenosis‘
## # i 25 more variables:
## #   ‘Decreased pulmonary capillary hydrostatic fluid pressure‘ <chr>,
## #   Oxytocin <chr>, ‘Graves’ disease‘ <chr>, B. <chr>,
## #   ‘Increased serum aldosterone concentration‘ <chr>, E....10 <chr>, ...
```

```r
unique(combined_df_new$Accomodations)
```

```
## [1] 2 1
```

```r
combined_df_new$Total_Score
```

```
##  [1] 0.70 0.70 0.70 0.70 0.70 0.70 0.60 0.60 0.60 0.60 0.60 0.55 0.50 0.50 0.50
## [16] 0.45 0.40 0.40 0.35 0.35 0.35 0.35 0.30 0.30 0.25 0.25 0.20 0.20 0.20 0.20
## [31] 0.15 0.40 0.25 0.50 0.75 0.35 0.65 0.45 0.55 0.45 0.30 0.75 0.10
```

```r
combined_df_new$Sex
```

```
##  [1] 0 1 0 1 1 0 1 1 1 1 1 0 0 1 1 1 1 0 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 0
## [39] 0 0 1 1 1
```

```r
combined_df_new %>%
  group_by(Sex) %>%                          # ← change to other factors too
  shapiro_test(Total_Score)
```

```
## # A tibble: 2 x 4
##      Sex variable    statistic      p
##    <dbl> <chr>           <dbl>  <dbl>
## 1      0 Total_Score     0.908 0.173
## 2      1 Total_Score     0.929 0.0469
```

```
combined_df_new$Total_Score
```

```
##  [1] 0.70 0.70 0.70 0.70 0.70 0.70 0.60 0.60 0.60 0.60 0.60 0.55 0.50 0.50 0.50
## [16] 0.45 0.40 0.40 0.35 0.35 0.35 0.35 0.30 0.30 0.25 0.25 0.20 0.20 0.20 0.20
## [31] 0.15 0.40 0.25 0.50 0.75 0.35 0.65 0.45 0.55 0.45 0.30 0.75 0.10
```

```
combined_df_new$Race_ethnicity
```

```
##  [1] 1 1 1 0 0 1 0 1 1 0 0 1 1 1 0 0 1 1 0 0 1 0 1 1 0 1 1 1 1 0 0 0 0 1 1 1 0 1
## [39] 1 1 0 1 1
```

```
combined_df_new %>%
  group_by(Race_ethnicity) %>%
  shapiro_test(Total_Score)
```

```
## # A tibble: 2 x 4
##   Race_ethnicity variable    statistic     p
##            <dbl> <chr>           <dbl> <dbl>
## 1              0 Total_Score     0.934 0.250
## 2              1 Total_Score     0.953 0.268
```

```
combined_df_new$Total_Score
```

```
##  [1] 0.70 0.70 0.70 0.70 0.70 0.70 0.60 0.60 0.60 0.60 0.60 0.55 0.50 0.50 0.50
## [16] 0.45 0.40 0.40 0.35 0.35 0.35 0.35 0.30 0.30 0.25 0.25 0.20 0.20 0.20 0.20
## [31] 0.15 0.40 0.25 0.50 0.75 0.35 0.65 0.45 0.55 0.45 0.30 0.75 0.10
```

```
combined_df_new$English_prof
```

```
##  [1] 0 1 1 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 1 0 1 1 1 1 0 0 1 0 1 0 1 0 0 1 0 1 0 0
## [39] 1 0 0 1 0
```

```
combined_df_new %>%
  group_by(English_prof) %>%
  shapiro_test(Total_Score)
```

```
## # A tibble: 2 x 4
##   English_prof variable    statistic     p
##          <dbl> <chr>           <dbl> <dbl>
## 1            0 Total_Score     0.952 0.262
## 2            1 Total_Score     0.937 0.282
```

```r
combined_df_new$Total_Score
```

```
##  [1] 0.70 0.70 0.70 0.70 0.70 0.70 0.60 0.60 0.60 0.60 0.60 0.55 0.50 0.50 0.50
## [16] 0.45 0.40 0.40 0.35 0.35 0.35 0.35 0.30 0.30 0.25 0.25 0.20 0.20 0.20 0.20
## [31] 0.15 0.40 0.25 0.50 0.75 0.35 0.65 0.45 0.55 0.45 0.30 0.75 0.10
```

```r
combined_df_new$Born_usa
```

```
##  [1] 0 0 1 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 1 0 0 0 1 0 0 0 1 1 0 1 0 0 0 0
## [39] 0 0 0 1 0
```

```r
combined_df_new %>%
  group_by(Born_usa) %>%
  shapiro_test(Total_Score)
```

```
## # A tibble: 2 x 4
##   Born_usa variable    statistic     p
##      <dbl> <chr>           <dbl> <dbl>
## 1        0 Total_Score     0.954 0.177
## 2        1 Total_Score     0.934 0.493
```

```r
combined_df_new$Total_Score
```

```
##  [1] 0.70 0.70 0.70 0.70 0.70 0.70 0.60 0.60 0.60 0.60 0.60 0.55 0.50 0.50 0.50
## [16] 0.45 0.40 0.40 0.35 0.35 0.35 0.35 0.30 0.30 0.25 0.25 0.20 0.20 0.20 0.20
## [31] 0.15 0.40 0.25 0.50 0.75 0.35 0.65 0.45 0.55 0.45 0.30 0.75 0.10
```

```r
combined_df_new$Home_language
```

```
##  [1] 0 1 1 0 1 0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 1 0 1 1 0 0 1 0 0 0 1 1 0 1 0 1 0 0
## [39] 1 0 0 1 0
```

```r
combined_df_new %>%
  group_by(Home_language) %>%
  shapiro_test(Total_Score)
```

```
## # A tibble: 2 x 4
##   Home_language variable    statistic     p
##           <dbl> <chr>           <dbl> <dbl>
## 1             0 Total_Score     0.948 0.204
## 2             1 Total_Score     0.953 0.506
```

```r
# Mann-Whitney U test for sex
wilcox.test(Total_Score ~ Sex, data = combined_df_new)
```

```
## Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Total_Score by Sex
## W = 260.5, p-value = 0.0843
## alternative hypothesis: true location shift is not equal to 0
```

```r
# Mann-Whitney U test for accommodations
wilcox.test(Total_Score ~ Accomodations, data = combined_df_new)
```

```
## Warning in wilcox.test.default(x = DATA[[1L]], y = DATA[[2L]], ...): cannot
## compute exact p-value with ties
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  Total_Score by Accomodations
## W = 40, p-value = 0.9769
## alternative hypothesis: true location shift is not equal to 0
```

```r
# t-tests for normal groups
t.test(Total_Score ~ Race_ethnicity, data = combined_df_new)
```

```
##
##  Welch Two Sample t-test
##
## data:  Total_Score by Race_ethnicity
## t = -0.4889, df = 35.874, p-value = 0.6279
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.14502910  0.08869426
## sample estimates:
## mean in group 0 mean in group 1
##       0.4352941       0.4634615
```

```r
t.test(Total_Score ~ English_prof, data = combined_df_new)
```

```
##
##  Welch Two Sample t-test
##
## data:  Total_Score by English_prof
## t = 0.3974, df = 33.963, p-value = 0.6936
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.09586908  0.14247541
## sample estimates:
## mean in group 0 mean in group 1
##       0.4615385       0.4382353
```

```r
t.test(Total_Score ~ Born_usa, data = combined_df_new)
```

```
##
##  Welch Two Sample t-test
##
## data:  Total_Score by Born_usa
## t = -0.046958, df = 13.158, p-value = 0.9633
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.1636148  0.1566451
## sample estimates:
## mean in group 0 mean in group 1
##       0.4515152       0.4550000
```

```
t.test(Total_Score ~ Home_language, data = combined_df_new)
```

```
##
##  Welch Two Sample t-test
##
## data:  Total_Score by Home_language
## t = -0.52079, df = 35.083, p-value = 0.6058
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -0.14793025  0.08752301
## sample estimates:
## mean in group 0 mean in group 1
##       0.4403846       0.4705882
```