STA234 HW7

Derin Gezgin

2025-04-29

100/100

Importing the Necessary Libraries

```
library(tidyverse)
library(pROC)
library(nycflights13)
```

Problem 1 [25 points]

Using the high school and beyond (hsb) data write a function to make comparative boxplots with variable1 as x, variable2 as y, and variable3 used to fill the boxes.

Set default as variable1 as `ses`, variable2 as `writing score`, and variable3 as `race`.

Note that: when you read the hsb data set all categorical variables as factors with correct categories so no matter which set of variables, we make the plots for the categories are correct.

Reading and preparation of the data

```
hsb = read.csv("hsb.csv")

hsb$female = factor(hsb$female,
                    levels = c(0, 1),
                    labels = c("Male",
                               "Female"))

hsb$ses = factor(hsb$ses,
                 levels = c(1, 2, 3),
                 labels = c("Low",
                            "Middle",
                            "High"))

hsb$race = factor(hsb$race,
                  levels = c(1, 2, 3, 4),
                  labels = c("Hispanic",
                             "Asian",
```

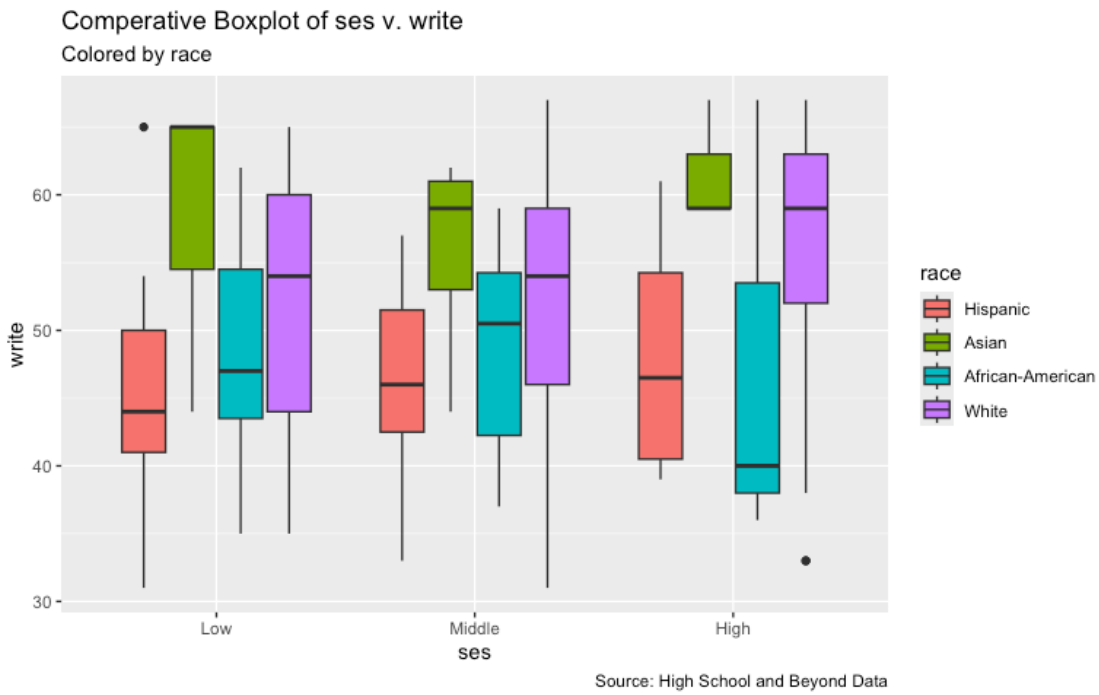
```
                              "African-American",
                              "White"))

hsb$schtyp = factor(hsb$schtyp,
                    levels = c(1, 2),
                    labels = c("Private",
                               "Public"))

hsb$prog = factor(hsb$prog,
                  levels = c(1, 2, 3),
                  labels = c("General",
                             "Academic",
                             "Vocational"))
```

Creating the function

```
hsbBoxPlotter = function(data = hsb,
                         x = "ses",
                         y = "write",
                         fill = "race") {
    ggplot(data = data,
           aes(x = .data[[x]],
               y = .data[[y]],
               fill = .data[[fill]])) +
        geom_boxplot() +
        labs(title = paste("Comperative Boxplot of", x, "v.", y),
             subtitle = paste("Colored by", fill),
             caption = paste("Source: High School and Beyond Data"))
}
```

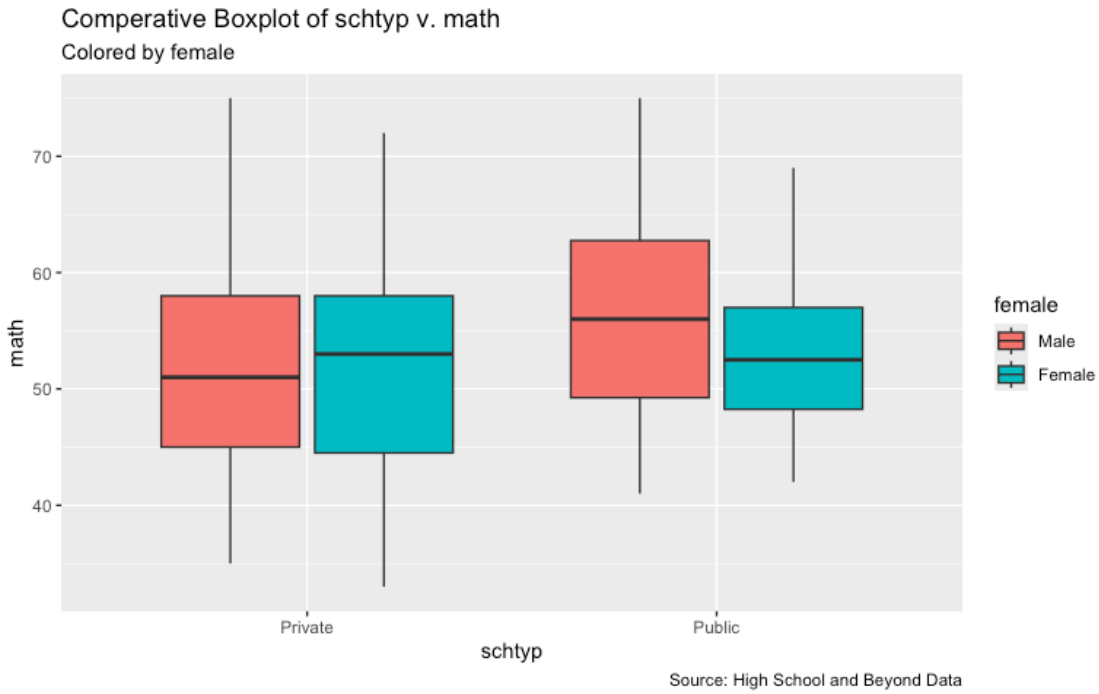
Testing with the default inputs

```
hsbBoxPlotter()
```



Testing with custom inputs

```
hsbBoxPlotter(x = "schtyp",
              y = "math",
              fill = "female")
```



Problem 2 [25 points]

Using the flights data from package nycflights13, write a function to group data by a given variable1 (input) and show minimum, maximum, and average of another variable2 (input).

Set default for variable1 as `carrier` and variable2 as `distance`.

```
flightSummary = function(variable1 = "carrier",
                         variable2 = "distance") {
    flights %>%
        group_by(.data[[variable1]]) %>%
        summarise(
            min = min(.data[[variable2]], na.rm = TRUE),
            max = max(.data[[variable2]], na.rm = TRUE),
            avg = mean(.data[[variable2]], na.rm = TRUE),
        )
}
```

Call function without any inputs and show output.

```
flightSummary()
```

```
## # A tibble: 16 × 4
##    carrier   min   max   avg
##    <chr>   <dbl> <dbl> <dbl>
##  1 9E         94  1587  530.
##  2 AA        187  2586 1340.
##  3 AS       2402  2402 2402
##  4 B6        173  2586 1069.
##  5 DL         94  2586 1237.
##  6 EV         80  1389  563.
##  7 F9       1620  1620 1620
##  8 FL        397   762  665.
##  9 HA       4983  4983 4983
## 10 MQ        184  1147  570.
## 11 OO        229  1008  501.
## 12 UA        116  4963 1529.
## 13 US         17  2153  553.
## 14 VX       2248  2586 2499.
## 15 WN        169  2133  996.
## 16 YV         96   544  375.
```

Call function with variable1 as `origin` and variable2 as `dep_delay` and show the output.

```
flightSummary(variable1 = "origin",
              variable2 = "dep_delay")
```

```
## # A tibble: 3 × 4
##   origin   min   max   avg
##   <chr>  <dbl> <dbl> <dbl>
## 1 EWR      -25  1126  15.1
## 2 JFK      -43  1301  12.1
## 3 LGA      -33   911  10.3
```

Problem 3 [25 points]

3.i

Write a for loop that prints the Displacement ('`disp`') of the '`mtcars`' dataset.

```
data(mtcars)
```

a. This loop will only print observations of 160 or higher in 'disp'.

```
for (disp in mtcars$disp) {
    if (disp >= 160) print(disp)
}
```

```
## [1] 160
## [1] 160
## [1] 258
## [1] 360
## [1] 225
## [1] 360
## [1] 167.6
## [1] 167.6
## [1] 275.8
## [1] 275.8
## [1] 275.8
## [1] 472
## [1] 460
## [1] 440
## [1] 318
## [1] 304
## [1] 350
## [1] 400
## [1] 351
## [1] 301
```

b. This loop will stop as soon as an observation is smaller than 160 in 'disp'.

```
for (disp in mtcars$disp) {
    if (disp < 160) break
    print(disp)
}
```

```
## [1] 160
## [1] 160
```

I am not sure if these should be in the same loop or not, so I created two separate versions of it. I am also adding a version with both conditions

```
for (disp in mtcars$disp) {
    if (disp < 160) { break }
    # Actually this condition is not needed
    # in the combined version as the program would
    # print regardless until there is something less than 160
    # Still added for better readibility
    else if (disp >= 160) { print(disp) }
}
```

```
## [1] 160
## [1] 160
```

3.ii

Write a while loop starting with `x = 0`. The loop prints all numbers up to 35 but it skips number 7.

```
x = 0
while (x <= 35) {
    if (x != 7) print(x)
    x = x + 1
}
```

```
## [1] 0
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 8
## [1] 9
## [1] 10
## [1] 11
## [1] 12
## [1] 13
## [1] 14
## [1] 15
## [1] 16
## [1] 17
## [1] 18
## [1] 19
## [1] 20
## [1] 21
## [1] 22
## [1] 23
## [1] 24
```

```
## [1] 25
## [1] 26
## [1] 27
## [1] 28
## [1] 29
## [1] 30
## [1] 31
## [1] 32
## [1] 33
## [1] 34
## [1] 35
```

3.iii

We are using the same while loop as in the last exercise. The loop prints again all numbers up to 35, but this time it skips a whole vector of numbers: 3,9,13,19,23,29. `exclude = c(3,9,13,19,23,29)`

```
exclude = c(3, 9, 13, 23, 29)
x = 0
while (x <= 35) {
    if (!(x %in% exclude)) print(x)
    x = x + 1
}
```

```
## [1] 0
## [1] 1
## [1] 2
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 10
## [1] 11
## [1] 12
## [1] 14
## [1] 15
## [1] 16
## [1] 17
## [1] 18
## [1] 19
## [1] 20
## [1] 21
## [1] 22
## [1] 24
## [1] 25
## [1] 26
## [1] 27
## [1] 28
```

```
## [1] 30
## [1] 31
## [1] 32
## [1] 33
## [1] 34
## [1] 35
```

3.iv

Use the ‘rivers’ dataset to write a for loop. The loop prints the dataset:

rivers shorter than 500 are a ‘short river’;

rivers longer than 2000 are a ‘long river’;

and rivers in the middle range are printed in their original numbers.

```
for (river in rivers) {
    if (river < 500) { print("short river") }
    else if (river > 2000) { print("long river") }
    else { print(river) }
}
```

```
## [1] 735
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] 524
## [1] "short river"
## [1] 1459
## [1] "short river"
## [1] "short river"
## [1] 600
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] 870
## [1] 906
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] 1000
## [1] 600
## [1] 505
## [1] 1450
## [1] 840
## [1] 1243
## [1] 890
## [1] "short river"
## [1] "short river"
## [1] "short river"
```

```
## [1] "short river"
## [1] 525
## [1] 720
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] 850
## [1] "short river"
## [1] 630
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] 730
## [1] 600
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] 710
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] 680
## [1] 570
## [1] "short river"
## [1] "short river"
## [1] 560
## [1] 900
## [1] 625
## [1] "short river"
## [1] "long river"
## [1] 1171
## [1] "long river"
## [1] "long river"
## [1] "long river"
## [1] 780
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] 760
```

```
## [1] 618
## [1] "short river"
## [1] 981
## [1] 1306
## [1] 500
## [1] 696
## [1] 605
## [1] "short river"
## [1] "short river"
## [1] 1054
## [1] 735
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] 1270
## [1] 545
## [1] "short river"
## [1] 1885
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] 800
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] 538
## [1] 1100
## [1] 1205
## [1] "short river"
## [1] "short river"
## [1] 610
## [1] "short river"
## [1] 540
## [1] 1038
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] "short river"
## [1] 620
## [1] "short river"
```

```
## [1] 652
## [1] 900
## [1] 525
## [1] "short river"
## [1] "short river"
## [1] 529
## [1] 500
## [1] 720
## [1] "short river"
## [1] "short river"
## [1] 671
## [1] 1770
```

Problem 4 [25 points]

Read the article WritingExample-2 in Moodle and report the following.

4.a

Explain the research goals of the study.

This study aims to investigate whether driving alone for long commutes (more than 30 minutes in one way) is associated with poor mental health outcomes at county level in the United States.

Specifically, the researchers are trying to determine

- Whether there is a statistically significant association between the percentage of the long commuters and the average number of poor mental health days reported by residents in the same county.
- Whether this association still holds true after other significant possible predictors such as social, economic, and health-related are taken into account.

4.b

Explain the dataset used in the study.

The researchers built their dataset from two public sources:

1. ***2020 County Health Rankings & Road-Maps Data***

   This dataset was obtained from the County Health Rankings & Roadmaps. It consists of various health outcome measures by US county, coming from sources like American Community Surveys, 5-year estimates, Behavioral Risk Factor Surveillance System, Bureau of Labor Statistics, using telephone surveys and other systems.

2. ***Missouri Census Data Center (MCDC) County Data***

   This dataset is used to calculate the population density of each county as it has the area of each county.

In the study, 21 ***explanatory variables*** are used which are found to be associated with mental health state. The researchers had two variables that identify the percentage of working population that drives alone to work, and the percentage of the population that drives alone to work for more than 30 minutes a day. A new variable is created to show the percentage of the entire work population that has a long (more than 30 minutes) commute.

The ***response variable*** is the average number of mentally unhealthy days reported in the 30 days.

The study also transformed several variables (*Unemployed, MedianIncome, and PopulationDensity)* using the log function to avoid the skewed distribution of variables. On the other hand, as *Suicides* and *DrugOverdoseDeaths* variables had missing values, which are ignored if missing, otherwise coded in a binary format.

In summary, the dataset contains data from 3,112 counties from all 50 states, each county being one observation.

4.c

Explain the methodology used to answer the research questions.

To figure out whether long solo commutes to work is associated to poor mental health, the researchers first built a basic model that included all the 21 predictor variables that has a possibility to be related to poor mental health, except the predictor of interest: the percentage of people driving alone for a long commute (LongCommuDriveAlone).

They then used the best subset analysis, which is basically trying out a bunch of different predictor variable combinations with four possible interactions (the interactions that have an interaction coefficient that is statistically significant at 1% level of significance) to find the combination that explains the mental health outcomes the best.

The full model is the same as the best subsets model, but it includes the LongCommuDriveAlone variable and its interactions.

Finally, to see if adding LongCommuDriveAlone actually made a meaningful difference, the study conducted an Extra Sum of Squares test which checks whether if this extra variable improves the model's ability to explain the poor mental health outcomes.

In summary, the researchers first constructed the best model without the predictor of interest followed by a model that includes that predictor, and finally they compared these two models to determine if the variable of interest is statistically significant or not.

4.d

Explain what is presented in the Appendix plots.

Appendix #1

This appendix shows the histogram of distribution of the 3 predictor variables: *Unemployed, MedianIncome, PopulationDensity*. From these histograms, we can see that the distribution of these variables are obviously skewed rather than showing a normal distribution, which requires a log transformation to spread out this skew.

Appendix #2

This correlation matrix shows the relationship between the predictor variables of the full model. For example the poor health condition is highly correlated with the food insecurity while suicide ratio and income has low correlation. This can be used to detect any multicollinearity between the variables of interest.

Appendix #3

This appendix shows a detailed summary of the final reduced regression model which includes parameters that are selected with the best-subsets analysis, with the statistically significant interaction terms. For each coefficient, the table shows its estimated values, the standard error, the t-statistic, and the p-value. At the same time, this output includes the model fit statistics like residual standard error, degrees of freedom, etc. From this summary, we can see that all predictors except food insecurity was significant in the model.

Appendix #4

This appendix shows the results of the Extra Sum of Squares (ESS) test that compares the final reduced model (that does not include the LongCommuDriveAlone and its interactions) with the full model (that includes them). As the description states, the low p-Value shows the significant difference in the goodness of fit of the model, which shows that adding the new variable and its interaction terms is statistically significant.

Appendix #5

In this set of plots which can be used to test the fit of the model.

The scatterplot of fitted values vs. the residuals checks the homoscedasticity assumption, and we cannot see a clear pan in/out pattern.

The Q-Q plot is used to check the normality assumption and we can see that the residuals are normally distributed.

The other residual plots can be used for us to check the linearity assumption, but we we would need further plots and tests to check this assumption.

Appendix #6

This shows a scatterplot of Long Commute Drive Alone variable against the Poor Mental Health Days variable. We cannot see any obvious patterns in this scatterplot which shows that Long Commute Drive Alone variable does not explain a significant portion of the variation in mental health days, without accounting for other variables.

Appendix #7

This appendix shows the summary of the full model with the complete regression variables, selected predictors, predictor of interest, and its interaction terms. In this summary we can find the coefficient estimate, standard error, t- and p-values. At the same time, we can see the residual standard error, multiple / adjusted R-squared, and the F-Statistic values. We can see the adjusted R-Squared value which is slightly higher than the reduced model, confirming that the predictor variable of interest and its interactions improves the explanatory power of the model.

Appendix #8

This group of plots show a group of plots which are colored by the Long Commute Drive variable, and the relationship between the other predictor variables and the response variable, Poor Mental Health Days. From these set of plots, we can see that the best-fit lines for the different categories of the Long Commute variable is similar which supports the claim of the paper. As the paper points out, the LongCommuDriveAlone variable improves the model, only when the other variables are taken into account.

Appendix #9

In this plot, we can see a map of the United States of America which has its counties colored by their average number of poor mental health days. Darker colors means less poor mental health days (the minimum in the scale is 3) while the lighter colors mean more mental health days (the maximum in the scale is 6). The paper points out that (we can also see this from the plot as well) some areas experience more poor mental health than others in general. which can lead the officials to do regional work.

4.e

Explain what would you do differently for the EDA to answer the research goals.

- I would definitely check for outliers in my data that can negatively affect my data analysis. These outliers can include small counties, as well as large counties.
- I can try to experiment with different kinds of transformations to transform the skewed variables in the dataset.
- While the paper groups the Drive Alone and Long Commute variables together, I would love to see how they individually interact with the data. It is possible that long