STATS 7022 - Data Science PG Assignment 1

Dang Thinh Nguyen

2024-06-24

1 Question 1: Data Analysis

```
# Read in the data
data <- readRDS('./board_game.rds')

# Display the first 10 lines of the data
head(data, 10)
```

```
## # A tibble: 10 x 32
##    num.x     id primary          description yearpublished minplayers maxplayers
##    <dbl>  <dbl> <chr>            <chr>                <dbl>      <dbl>      <dbl>
##  1     0  30549 Pandemic         In Pandemi~           2008          2          4
##  2     1    822 Carcassonne      Carcassonn~           2000          2          5
##  3     2     13 Catan            In CATAN (~           1995          3          4
##  4     3  68448 7 Wonders        You are th~           2010          2          7
##  5     4  36218 Dominion         &quot;You ~           2008          2          4
##  6     5   9209 Ticket to Ride   With elega~           2004          2          5
##  7     6 178900 Codenames        Codenames ~           2015          2          8
##  8     7 167791 Terraforming Ma~ In the 240~           2016          1          5
##  9     8 173346 7 Wonders Duel   In many wa~           2015          2          2
## 10     9  31260 Agricola         Descriptio~           2007          1          5
## # i 25 more variables: playingtime <dbl>, minplaytime <dbl>, maxplaytime <dbl>,
## #   minage <dbl>, boardgamecategory <chr>, boardgamemechanic <chr>,
## #   boardgamefamily <chr>, boardgameexpansion <chr>,
## #   boardgameimplementation <chr>, boardgamedesigner <chr>,
## #   boardgameartist <chr>, boardgamepublisher <chr>, owned <dbl>,
## #   trading <dbl>, wanting <dbl>, wishing <dbl>, num.y <dbl>, name <chr>,
## #   year <dbl>, rank <dbl>, average <dbl>, bayes_average <dbl>, ...
```

1.1 (a). Select variables

```
columns <- c('primary', 'year', 'boardgamemechanic',
             'minplaytime', 'maxplaytime', 'average')
data2 <- data %>% select(columns)

# Display the first 10 lines of the data
head(data2, 10)
```

```
## # A tibble: 10 x 6
##    primary            year boardgamemechanic     minplaytime maxplaytime average
##    <chr>             <dbl> <chr>                       <dbl>       <dbl>   <dbl>
##  1 Pandemic           2008 ['Action Points', 'C~          45          45    7.59
##  2 Carcassonne        2000 ['Area Majority / In~          30          45    7.42
##  3 Catan              1995 ['Dice Rolling', 'He~          60         120    7.14
##  4 7 Wonders          2010 ['Drafting', 'Hand M~          30          30    7.74
##  5 Dominion           2008 ['Deck, Bag, and Poo~          30          30    7.61
##  6 Ticket to Ride     2004 ['Card Drafting', 'C~          30          60    7.41
##  7 Codenames          2015 ['Communication Limi~          15          15    7.6
##  8 Terraforming Mars  2016 ['Drafting', 'End Ga~         120         120    8.42
##  9 7 Wonders Duel     2015 ['Card Drafting', 'L~          30          30    8.11
## 10 Agricola           2007 ['Advantage Token', ~          30         150    7.93
```

```
# Display the column names
colnames(data2)
```

```
## [1] "primary"           "year"              "boardgamemechanic"
## [4] "minplaytime"       "maxplaytime"       "average"
```

1.2 (b). Rename average to rating

```
data3 <- data2 %>% rename(rating = average)

# Display the first 10 lines of the data
head(data3, 10)
```

```
## # A tibble: 10 x 6
##    primary            year boardgamemechanic      minplaytime maxplaytime rating
##    <chr>             <dbl> <chr>                        <dbl>       <dbl>  <dbl>
##  1 Pandemic           2008 ['Action Points', 'Co~          45          45   7.59
##  2 Carcassonne        2000 ['Area Majority / Inf~          30          45   7.42
##  3 Catan              1995 ['Dice Rolling', 'Hex~          60         120   7.14
##  4 7 Wonders          2010 ['Drafting', 'Hand Ma~          30          30   7.74
##  5 Dominion           2008 ['Deck, Bag, and Pool~          30          30   7.61
##  6 Ticket to Ride     2004 ['Card Drafting', 'Co~          30          60   7.41
##  7 Codenames          2015 ['Communication Limit~          15          15   7.6
##  8 Terraforming Mars  2016 ['Drafting', 'End Gam~         120         120   8.42
##  9 7 Wonders Duel     2015 ['Card Drafting', 'La~          30          30   8.11
## 10 Agricola           2007 ['Advantage Token', '~          30         150   7.93
```

```
colnames(data3)
```

```
## [1] "primary"           "year"              "boardgamemechanic"
## [4] "minplaytime"       "maxplaytime"       "rating"
```

1.3 (c). Remove any games released before 2016 or after 2020

```
data4 <- data3 %>%
  filter((year >= 2016) & (year <= 2020))

# Display the first 10 lines of the data
head(data4, 10)
```

```
## # A tibble: 10 x 6
##    primary                 year boardgamemechanic minplaytime maxplaytime rating
##    <chr>                  <dbl> <chr>                    <dbl>       <dbl>  <dbl>
##  1 Terraforming Mars       2016 ['Drafting', 'En~          120         120   8.42
##  2 Scythe                  2016 ['Action Draftin~           90         115   8.22
##  3 Azul                    2017 ['End Game Bonus~           30          45   7.8
##  4 Wingspan                2019 ['Card Drafting'~           40          70   8.1
##  5 Gloomhaven              2017 ['Action Queue',~           60         120   8.74
##  6 Kingdomino              2016 ['Card Drafting'~           15          15   7.35
##  7 Arkham Horror: The Ca~  2016 ['Action Points'~           60         120   8.16
##  8 Great Western Trail     2016 ['Deck, Bag, and~           75         150   8.29
##  9 Spirit Island           2017 ['Action Retriev~           90         120   8.36
## 10 Clank!: A Deck-Buildi~  2016 ['Card Drafting'~           30          60   7.82
```

```
# Display the summary of the year variable
summary(data4$year)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2016    2017    2018    2018    2019    2020
```

1.4 (d). Create duration variable and remove any games with zero duration

```
# Create duration variable
data5 <- data4 %>%
  mutate(duration = sqrt(minplaytime * maxplaytime))

# Remove zero duration
data5 <- data5 %>%
  filter(duration != 0)

# Display the first 10 lines of the data
head(data5, 10)
```

```
## # A tibble: 10 x 7
##    primary        year boardgamemechanic minplaytime maxplaytime rating duration
##    <chr>         <dbl> <chr>                    <dbl>       <dbl>  <dbl>    <dbl>
##  1 Terraforming~  2016 ['Drafting', 'En~          120         120   8.42    120
##  2 Scythe         2016 ['Action Draftin~           90         115   8.22    102.
##  3 Azul           2017 ['End Game Bonus~           30          45   7.8      36.7
```

```
##  4 Wingspan       2019 ['Card Drafting'~          40          70   8.1      52.9
##  5 Gloomhaven     2017 ['Action Queue',~          60         120   8.74     84.9
##  6 Kingdomino     2016 ['Card Drafting'~          15          15   7.35     15
##  7 Arkham Horro~  2016 ['Action Points'~          60         120   8.16     84.9
##  8 Great Wester~  2016 ['Deck, Bag, and~          75         150   8.29    106.
##  9 Spirit Island  2017 ['Action Retriev~          90         120   8.36    104.
## 10 Clank!: A De~  2016 ['Card Drafting'~          30          60   7.82     42.4
```

```
# Display the summary of duration variable
summary(data5$duration)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   21.21   36.74   52.35   60.00 7745.97
```

1.5 (e). Create dice variable

```
data6 <- data5 %>%
  mutate(
    dice = ifelse(str_detect(boardgamemechanic, 'Dice'), 'Dice game', 'Not a dice game'),
    dice = replace_na(dice, 'Not a dice game')
  )

# Display the first 10 lines of the data
head(data6, 10)
```

```
## # A tibble: 10 x 8
##    primary  year boardgamemechanic minplaytime maxplaytime rating duration dice
##    <chr>   <dbl> <chr>                   <dbl>       <dbl>  <dbl>    <dbl> <chr>
##  1 Terraf~  2016 ['Drafting', 'En~         120         120   8.42    120   Not ~
##  2 Scythe   2016 ['Action Draftin~          90         115   8.22    102.  Not ~
##  3 Azul     2017 ['End Game Bonus~          30          45   7.8      36.7 Not ~
##  4 Wingsp~  2019 ['Card Drafting'~          40          70   8.1      52.9 Dice~
##  5 Gloomh~  2017 ['Action Queue',~          60         120   8.74     84.9 Not ~
##  6 Kingdo~  2016 ['Card Drafting'~          15          15   7.35     15   Not ~
##  7 Arkham~  2016 ['Action Points'~          60         120   8.16     84.9 Not ~
##  8 Great ~  2016 ['Deck, Bag, and~          75         150   8.29    106.  Not ~
##  9 Spirit~  2017 ['Action Retriev~          90         120   8.36    104.  Not ~
## 10 Clank!~  2016 ['Card Drafting'~          30          60   7.82     42.4 Not ~
```

```
# Display the count of dice observations
data6 %>% count(dice)
```

```
## # A tibble: 2 x 2
##   dice               n
##   <chr>          <int>
## 1 Dice game       1587
## 2 Not a dice game  4250
```

1.6 (f). Create rating_7.5 variable

```
data7 <- data6 %>%
  mutate(rating_7.5 = ifelse(rating >= 7.5, TRUE, FALSE))

# Display the first 10 lines of the data
head(data7, 10)
```

```
## # A tibble: 10 x 9
##    primary  year boardgamemechanic minplaytime maxplaytime rating duration dice
##    <chr>   <dbl> <chr>                   <dbl>       <dbl>  <dbl>    <dbl> <chr>
##  1 Terraf~  2016 ['Drafting', 'En~         120         120   8.42    120   Not ~
##  2 Scythe   2016 ['Action Draftin~          90         115   8.22    102.  Not ~
##  3 Azul     2017 ['End Game Bonus~          30          45   7.8      36.7 Not ~
##  4 Wingsp~  2019 ['Card Drafting'~          40          70   8.1      52.9 Dice~
##  5 Gloomh~  2017 ['Action Queue',~          60         120   8.74     84.9 Not ~
##  6 Kingdo~  2016 ['Card Drafting'~          15          15   7.35     15   Not ~
##  7 Arkham~  2016 ['Action Points'~          60         120   8.16     84.9 Not ~
##  8 Great ~  2016 ['Deck, Bag, and~          75         150   8.29    106.  Not ~
##  9 Spirit~  2017 ['Action Retriev~          90         120   8.36    104.  Not ~
## 10 Clank!~  2016 ['Card Drafting'~          30          60   7.82     42.4 Not ~
## # i 1 more variable: rating_7.5 <lgl>
```

```
# Display the count of rating_7.5 observations
data7 %>% count(rating_7.5)
```

```
## # A tibble: 2 x 2
##   rating_7.5     n
##   <lgl>      <int>
## 1 FALSE       4675
## 2 TRUE        1162
```

1.7 (g). Produce boxplots

```
data7 %>%
  ggplot(aes(x = factor(year), y = log(duration), fill = rating_7.5)) +
  geom_boxplot(outlier.shape = 19) +
  facet_wrap(. ~ dice, nrow = 2) +
  labs(title = 'Game Duration by Year',
       subtitle = 'Colour indicates if game has a rating of 7.5 or above.',
       x = 'Year',
       y = 'Duration (minutes)',
       fill = expression('Rating' >= 7.5),
       caption = 'Note: Games without a mechanic have been classified as "Not a dice game".') +
  scale_y_continuous(
    breaks = log(c(1, 10, 100, 1000, 10000)),
    labels = c(1, 10, 100, 1000, 10000)
```

```
) +
scale_fill_manual(values = c("FALSE" = "#FFDF22", "TRUE" = "#0C8990")) +
theme_minimal() +
theme_bw() +
theme(
  #panel.border = element_rect(color = "black", fill = NA, linewidth = 0.3),
  #strip.background = element_rect(color = "black", fill = 'lightgray', linewidth = 0.3),
  legend.position = 'bottom'
)
```

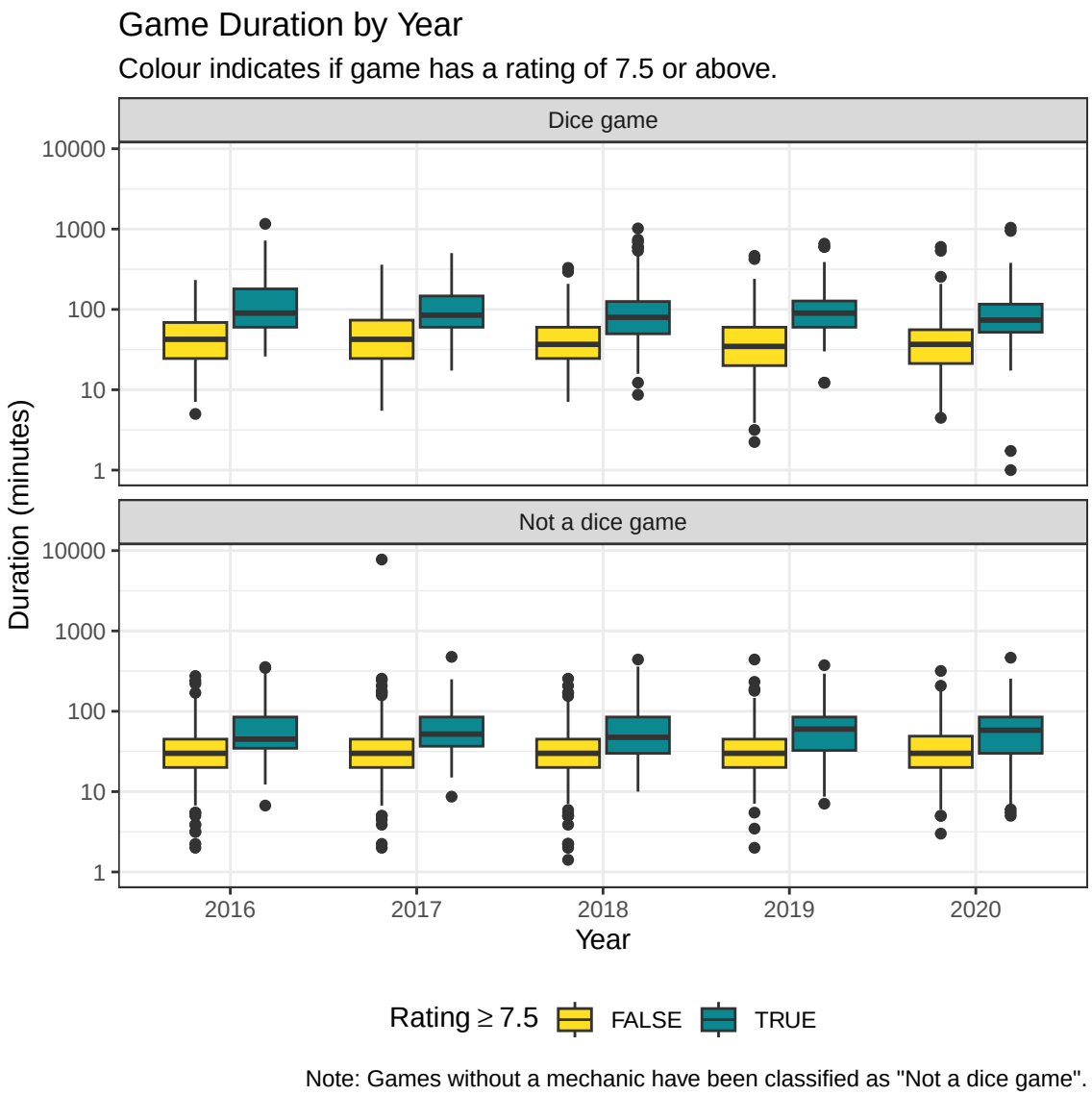

Figure 1: Side-by-side boxplot of game duration (in minutes) by year, separated by whether or not each game is a dice game, and whether or not each game has a rating of at least 7.5.