Star Wars Reproduction

Armelle Duston

2025-01-21

Setup

In this first step, we install (if necessary) and import relevant libraries.

```
# Vector of package names
packages <- c("ggplot2", "dplyr")

# Install any packages not yet installed
installed_packages <- packages %in% rownames(installed.packages())
if (any(installed_packages == FALSE)) {
  install.packages(packages[!installed_packages])
}

# Load packages
invisible(lapply(packages, library, character.only = TRUE))
```

```
##
## 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
```

Import Data

Import data from "StarWars.csv". Raw data has column names on two different rows, manually define column-names and remove spaces for ease of use.

```
df <- read.csv("StarWars.csv", skip = 2, header = FALSE)

colnames(df) <- c("RespondentID",
                  "Seen_any",
                  "StarWars_fan",
                  "Seen_Ep_I",
                  "Seen_Ep_II",
```

```
                  "Seen_Ep_III",
                  "Seen_Ep_IV",
                  "Seen_Ep_V",
                  "Seen_Ep_VI",
                  "Rank_Ep_I",
                  "Rank_Ep_II",
                  "Rank_Ep_III",
                  "Rank_Ep_IV",
                  "Rank_Ep_V",
                  "Rank_Ep_VI",
                  "HanSolo",
                  "LukeSkywalker",
                  "PrincessLeiaOrgana",
                  "AnakinSkywalker",
                  "ObiWanKenobi",
                  "EmperorPalpatine",
                  "DarthVader",
                  "LandoCalrissian",
                  "BobaFett",
                  "C-3PO",
                  "R2D2",
                  "JarJarBinks",
                  "PadmeAmidala",
                  "Yoda",
                  "Which_character_shot_first",
                  "Familiar_with_Expanded_Universe",
                  "Expanded_Universe_fan",
                  "StarTrek_fan",
                  "Gender",
                  "Age",
                  "Household_Income",
                  "Education",
                  "Location")
```

Preprocess Data

Code the columns below as 0s and 1s for simplicity.

```
cols_to_convert <- c("Seen_Ep_I",
                  "Seen_Ep_II",
                  "Seen_Ep_III",
                  "Seen_Ep_IV",
                  "Seen_Ep_V",
                  "Seen_Ep_VI")

df[cols_to_convert] <- lapply(df[cols_to_convert], function(x) { as.integer(x != "") })
```

Figure 1

Recreate Figure 1: “Which ‘Star Wars’ Movies Have You Seen?”

```r
# Filter data for respondents that have seen at least one of the films
filtered_df1 <- filter(df, Seen_Ep_I == 1 |
                         Seen_Ep_II == 1 |
                         Seen_Ep_III == 1 |
                         Seen_Ep_IV == 1 |
                         Seen_Ep_V == 1 |
                         Seen_Ep_VI == 1)


# Make a dataframe to plot the barchart
fig1_df <- data.frame(round(apply(filtered_df1[,4:9],
                            MARGIN = 2,
                            FUN = sum)/length(filtered_df1$Seen_any)*100),
                      row.names = NULL)
colnames(fig1_df) <- "pct_seen"
fig1_df$film_names <- c("The Phantom Menace",
                        "Attack of the Clones",
                        "Revenge of the Sith",
                        "A New Hope",
                        "The Empire Strikes Back",
                        "Return of the Jedi")


# Use ggplot to recreate figure 1
ggplot(fig1_df, aes(x = film_names, y = pct_seen)) +
  geom_col(width = 0.6, fill = "#008FD5") +
  coord_flip() +
  geom_text(aes(label = paste0(pct_seen, "%")),
            hjust = -0.2,
            color = "black",
            size = 5) +
  scale_y_continuous(limits = c(0, 100), expand = c(0, 0)) +
  labs(
    x = NULL,
    y = NULL,
    title = "Which 'Star Wars' Movies Have You Seen?",
    subtitle = paste("Of", length(filtered_df1$Seen_any), "respondents who have seen any film"),
    caption = "Source: SurveyMonkey Audience"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    panel.grid.major.y = element_blank(),
    panel.grid.minor = element_blank(),

    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(size = 13, margin = margin(b = 15)),
    plot.caption = element_text(size = 10, hjust = 1,
                                margin = margin(t = 15)),

    axis.text.x = element_blank(),
    axis.ticks = element_blank() )
```

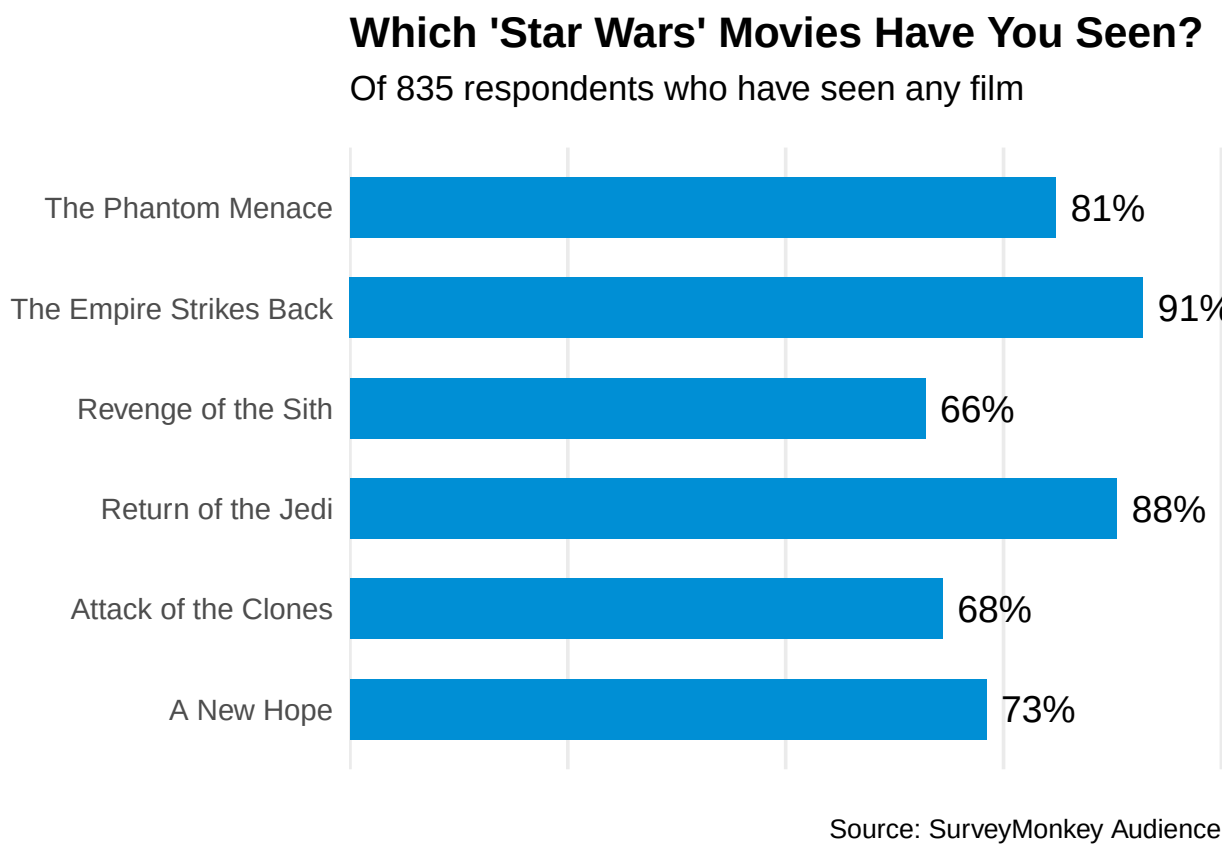
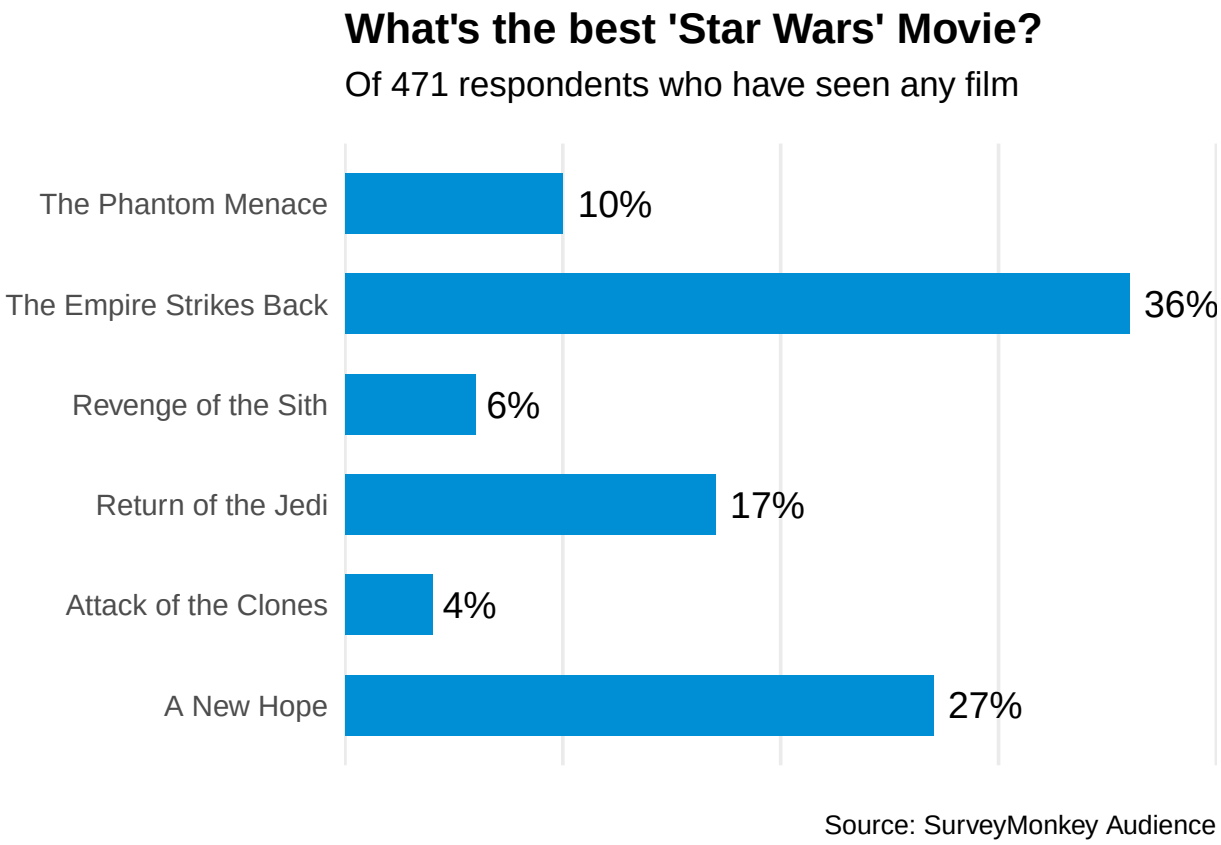


Figure 2

Recreate Figure 2: "What's the best 'Star Wars' Movie?"

```
# Filter data for respondents that have seen all of the films
filtered_df2 <- filter(df, Seen_Ep_I == 1 &
                         Seen_Ep_II == 1 &
                         Seen_Ep_III == 1 &
                         Seen_Ep_IV == 1 &
                         Seen_Ep_V == 1 &
                         Seen_Ep_VI == 1)


# Observation number 177 had NA for film 3 ranking, filled in based on other rankings
filtered_df2$Rank_Ep_III[177] <- 6


# Make a dataframe to plot the barchart
fig2_df <- data.frame(round(apply(filtered_df2[,10:15], 2,
                       function(col) sum(col == 1))/length(filtered_df2$Seen_any)*100),
                      row.names = NULL)
colnames(fig2_df) <- "ranked_1"
fig2_df$film_names <- c("The Phantom Menace",
                        "Attack of the Clones",
                        "Revenge of the Sith",
```

```
                        "A New Hope",
                        "The Empire Strikes Back",
                        "Return of the Jedi")


# Use ggplot to recreate figure 2
ggplot(fig2_df, aes(x = film_names, y = ranked_1)) +
  geom_col(width = 0.6, fill = "#008FD5") +
  coord_flip() +
  geom_text(aes(label = paste0(ranked_1, "%")),
            hjust = -0.2,
            color = "black",
            size = 5) +
  scale_y_continuous(limits = c(0, 40), expand = c(0, 0)) +
  labs(
    x = NULL,
    y = NULL,
    title = "What's the best 'Star Wars' Movie?",
    subtitle = paste("Of", length(filtered_df2$Seen_any), "respondents who have seen any film"),
    caption = "Source: SurveyMonkey Audience"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    panel.grid.major.y = element_blank(),
    panel.grid.minor = element_blank(),

    plot.title = element_text(face = "bold", size = 16),
    plot.subtitle = element_text(size = 13, margin = margin(b = 15)),
    plot.caption = element_text(size = 10, hjust = 1,
                                margin = margin(t = 15)),

    axis.text.x = element_blank(),
    axis.ticks = element_blank() )
```



Source: SurveyMonkey Audience