

FinalData

Import Dataset

```
library(tidyverse)
```

```
-- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
v dplyr      1.1.3      v readr      2.1.4
v forcats    1.0.0      v stringr    1.5.0
v ggplot2    3.4.3      v tibble     3.2.1
v lubridate  1.9.2      v tidyr      1.3.0
v purrr      1.0.2

-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()     masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become
```

```
library(skimr)
library(readr)
library(utf8)

# Check the encoding of the CSV file
file_encoding <- guess_encoding("week7_starwars.csv")$encoding
cat("File Encoding:", file_encoding, "\n")
```

File Encoding: windows-1252

```
# Read CSV file and convert strings to UTF-8 encoding
starwars <- read_csv("week7_starwars.csv", locale = locale(encoding = file_encoding)) %>%
  mutate(across(where(is.character), utf8_encode))
```

```

New names:
Rows: 1187 Columns: 38
-- Column specification
----- Delimiter: "," chr
(37): Have you seen any of the 6 films in the Star Wars franchise?, Do y... dbl
(1): RespondentID
i Use `spec()` to retrieve the full column specification for this data. i
Specify the column types or set `show_col_types = FALSE` to quiet this message.
* `` -> `...5`
* `` -> `...6`
* `` -> `...7`
* `` -> `...8`
* `` -> `...9`
* `` -> `...11`
* `` -> `...12`
* `` -> `...13`
* `` -> `...14`
* `` -> `...15`
* `` -> `...17`
* `` -> `...18`
* `` -> `...19`
* `` -> `...20`
* `` -> `...21`
* `` -> `...22`
* `` -> `...23`
* `` -> `...24`
* `` -> `...25`
* `` -> `...26`
* `` -> `...27`
* `` -> `...28`
* `` -> `...29`

```

```

# Use skim on the dataframe
skim(starwars)

```

Table 1: Data summary

Name	starwars
Number of rows	1187
Number of columns	38
Column type frequency:	

character	37
numeric	1
<hr/>	
Group variables	None
<hr/>	

Variable type: character

skim_variable	n	missing	complete	na	empty	unique	whitespace
Have you seen any of the 6 films in the Star Wars franchise?	0	1.00	2	8	0	3	0
Do you consider yourself to be a fan of the Star Wars film franchise?	350	0.71	2	8	0	3	0
Which of the following Star Wars films have you seen? Please select all that apply.	513	0.57	40	40	0	1	0
...5	615	0.48	43	43	0	1	0
...6	636	0.46	43	43	0	1	0
...7	579	0.51	33	33	0	1	0
...8	428	0.64	44	44	0	1	0
...9	448	0.62	40	40	0	1	0
Please rank the Star Wars films in order of preference with 1 being your favorite film in the franchise and 6 being your least favorite film.	351	0.70	1	40	0	7	0
...11	350	0.71	1	43	0	7	0
...12	351	0.70	1	43	0	7	0
...13	350	0.71	1	33	0	7	0
...14	350	0.71	1	44	0	7	0
...15	350	0.71	1	40	0	7	0
Please state whether you view the following characters favorably, unfavorably, or are unfamiliar with him/her.	357	0.70	8	43	0	7	0
...17	355	0.70	14	43	0	7	0
...18	355	0.70	14	43	0	7	0
...19	363	0.69	14	43	0	7	0
...20	361	0.70	14	43	0	7	0
...21	372	0.69	14	43	0	7	0
...22	360	0.70	11	43	0	7	0
...23	366	0.69	14	43	0	7	0
...24	374	0.68	9	43	0	7	0
...25	359	0.70	5	43	0	7	0
...26	356	0.70	5	43	0	7	0

skim_variable	n_missing	complete_rate	min	max	empty	unique	whitespace
...27	365	0.69	13	43	0	7	0
...28	372	0.69	13	43	0	7	0
...29	360	0.70	4	43	0	7	0
Which character shot first?	358	0.70	3	32	0	4	0
Are you familiar with the Expanded Universe?	358	0.70	2	8	0	3	0
Do you consider yourself to be a fan of the Expanded Universe? (Eæ	973	0.18	2	8	0	3	0
Do you consider yourself to be a fan of the Star Trek franchise?	118	0.90	2	8	0	3	0
Gender	140	0.88	4	8	0	3	0
Age	140	0.88	4	8	0	5	0
Household Income	328	0.72	8	19	0	6	0
Education	150	0.87	8	32	0	6	0
Location (Census Region)	143	0.88	7	18	0	10	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
RespondentID	1	329012820055633	2883729322894509629	0014713729008144329	2879998					

I don't think pivoting is super necessary, since every row is a separate person who responded.

The variable types in each column seem to be good. Most are char.

I need to remove some NAs using R code.

Fixing column names are necessary because they are full questions with spaces, so I need to shorten them and remove spaces.