Project title

Proposal

Brilliant Cassowary

```r
library(tidyverse)
library(skimr)
```

Data 1: Most popular coffee or beverage

Introduction and data

- Identify the source of the data.

  The data was sourced from Data Is Plural, a newsletter that curates interesting datasets. (actual link: https://bit.ly/gacttCSV+)

- State when and how it was originally collected (by the original data curator, not necessarily how you found the data).

  The original data was collected through a comprehensive survey conducted by British Youtuber and former World Barista Champion James Hoffman during the Great American Coffee Taste Test. He published this anonymized data and a video summarizing the results. The survey aimed to gather insights into coffee consumption habits, preferences, and demographic information of coffee drinkers.

- Write a brief description of the observations.

  This dataset contains a wide array of information related to coffee consumption, including the number of cups consumed per day, preferred locations for drinking coffee, and demographic details such as age, gender, education level, employment status, and more. The data provides a detailed snapshot of coffee drinking habits across different population segments.

Research question

- A well formulated research question. (You may include more than one research question if you want to receive feedback on different ideas for your project. However, one per data set is required.)
- A description of the research topic along with a concise statement of your hypotheses on this topic.
- Identify the types of variables in your research question. Categorical? Quantitative?

    1. Research Question: How do daily coffee consumption habits vary across different age groups?

    - This study aims to investigate the patterns of daily coffee consumption among various age groups. The focus is on understanding how coffee drinking habits differ across different demographics, specifically looking at the average number of cups consumed per day.

      Hypothesis: Young adults aged 18-24 consume more cups of coffee per day compared to other age groups.

      Variables: Categorical: Age group Quantitative: Number of cups of coffee consumed per day

    2. Research Question: What are the preferred locations for drinking coffee among different demographic groups?

    - This study aims to explore the preferred locations for drinking coffee among different demographic groups. The focus is on identifying patterns in coffee consumption behavior, particularly where individuals choose to enjoy their coffee, such as at home, in coffee shops, at work, or other locations.

      Hypothesis: The majority of students prefer to drink coffee at home rather than at coffee shops.

      Variables: Categorical: Preferred coffee drinking location, demographic group (e.g., students, employed individuals)

Glimpse of data

```
library(skimr)
library(readr)
read_csv("data/GACTT_RESULTS_ANONYMIZED_v2.csv") |>
  skim()
```

```
Rows: 4042 Columns: 113
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (44): Submission ID, What is your age?, How many cups of coffee do you t...
dbl (13): Lastly, how would you rate your own coffee expertise?, Coffee A - ...
lgl (56): Where do you typically drink coffee? (At home), Where do you typic...

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.
```

Table 1: Data summary

Name	read_csv("data/GACTT_RESU…
Number of rows	4042
Number of columns	113
Column type frequency:	
character	44
logical	56
numeric	13
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Submission ID	0	1.00	6	6	0	4042	0
What is your age?	31	0.99	13	15	0	7	0
How many cups of coffee do you typically drink per day?	93	0.98	1	11	0	6	0
Where do you typically drink coffee?	70	0.98	7	44	0	65	0
How do you brew coffee at home?	385	0.90	5	165	0	449	0
How else do you brew coffee at home?	3364	0.17	2	319	0	160	0
On the go, where do you typically purchase coffee?	3332	0.18	5	107	0	89	0
Where else do you purchase coffee?	4011	0.01	4	83	0	26	0
What is your favorite coffee drink?	62	0.98	5	32	0	12	0
Please specify what your favorite coffee drink is	3926	0.03	3	92	0	78	0
Do you usually add anything to your coffee?	83	0.98	5	100	0	53	0
What else do you add to your coffee?	3994	0.01	3	140	0	42	0
What kind of dairy do you add?	2356	0.42	8	110	0	175	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
What kind of sugar or sweetener do you add?	3530	0.13	5	99	0	82	0
Before today's tasting, which of the following best described what kind of coffee you like?	84	0.98	4	11	0	12	0
How strong do you like your coffee?	126	0.97	4	15	0	5	0
What roast level of coffee do you prefer?	102	0.97	4	7	0	7	0
How much caffeine do you like in your coffee?	125	0.97	5	13	0	3	0
Coffee A - Notes	1464	0.64	1	377	0	2317	0
Coffee B - Notes	1586	0.61	1	980	0	2199	0
Coffee C - Notes	1659	0.59	1	438	0	2163	0
Coffee D - Notes	1454	0.64	1	528	0	2354	0
Between Coffee A, Coffee B, and Coffee C which did you prefer?	270	0.93	8	8	0	3	0
Between Coffee A and Coffee D, which did you prefer?	281	0.93	8	8	0	2	0
Lastly, what was your favorite overall coffee?	272	0.93	8	8	0	4	0
Do you work from home or in person?	518	0.87	18	26	0	3	0
In total, much money do you typically spend on coffee in a month?	531	0.87	4	8	0	6	0
Why do you drink coffee?	474	0.88	5	93	0	84	0
Other reason for drinking coffee	3875	0.04	2	195	0	163	0
Do you like the taste of coffee?	479	0.88	2	3	0	2	0
Do you know where your coffee comes from?	483	0.88	2	3	0	2	0
What is the most you've ever paid for a cup of coffee?	515	0.87	5	13	0	8	0
What is the most you'd ever be willing to pay for a cup of coffee?	532	0.87	5	13	0	8	0
Do you feel like you're getting good value for your money when you buy coffee at a cafe?	542	0.87	2	3	0	2	0
Approximately how much have you spent on coffee equipment in the past 5 years?	536	0.87	7	16	0	7	0
Do you feel like you're getting good value for your money with regards to your coffee equipment?	548	0.86	2	3	0	2	0
Gender	519	0.87	4	22	0	5	0
Gender (please specify)	4030	0.00	2	28	0	11	0
Education Level	604	0.85	15	34	0	6	0
Ethnicity/Race	624	0.85	15	29	0	6	0
Ethnicity/Race (please specify)	3937	0.03	2	53	0	82	0
Employment Status	623	0.85	7	18	0	6	0
Number of Children	636	0.84	1	11	0	5	0
Political Affiliation	753	0.81	8	14	0	4	0

Variable type: logical

skim_variable	n_missing	complete_rate	mean	count
Where do you typically drink coffee? (At home)	67	0.98	0.92	TRU: 3644, FAL: 331
Where do you typically drink coffee? (At the office)	67	0.98	0.36	FAL: 2545, TRU: 1430
Where do you typically drink coffee? (On the go)	67	0.98	0.18	FAL: 3270, TRU: 705
Where do you typically drink coffee? (At a cafe)	67	0.98	0.29	FAL: 2805, TRU: 1170
Where do you typically drink coffee? (None of these)	67	0.98	0.01	FAL: 3939, TRU: 36
How do you brew coffee at home? (Pour over)	381	0.91	0.63	TRU: 2295, FAL: 1366
How do you brew coffee at home? (French press)	381	0.91	0.20	FAL: 2926, TRU: 735
How do you brew coffee at home? (Espresso)	381	0.91	0.41	FAL: 2143, TRU: 1518
How do you brew coffee at home? (Coffee brewing machine (e.g. Mr. Coffee))	381	0.91	0.18	FAL: 2998, TRU: 663
How do you brew coffee at home? (Pod/capsule machine (e.g. Keurig/Nespresso))	381	0.91	0.09	FAL: 3325, TRU: 336
How do you brew coffee at home? (Instant coffee)	381	0.91	0.04	FAL: 3531, TRU: 130
How do you brew coffee at home? (Bean-to-cup machine)	381	0.91	0.02	FAL: 3577, TRU: 84
How do you brew coffee at home? (Cold brew)	381	0.91	0.14	FAL: 3136, TRU: 525
How do you brew coffee at home? (Coffee extract (e.g. Cometeer))	381	0.91	0.05	FAL: 3475, TRU: 186
How do you brew coffee at home? (Other)	381	0.91	0.18	FAL: 2984, TRU: 677
On the go, where do you typically purchase coffee? (National chain (e.g. Starbucks, Dunkin))	3319	0.18	0.46	FAL: 394, TRU: 329
On the go, where do you typically purchase coffee? (Local cafe)	3319	0.18	0.54	TRU: 392, FAL: 331
On the go, where do you typically purchase coffee? (Drive-thru)	3319	0.18	0.13	FAL: 628, TRU: 95
On the go, where do you typically purchase coffee? (Specialty coffee shop)	3319	0.18	0.61	TRU: 438, FAL: 285

skim_variable	n_missing	complete_rate	mean	count
On the go, where do you typically purchase coffee? (Deli or supermarket)	3319	0.18	0.07	FAL: 674, TRU: 49
On the go, where do you typically purchase coffee? (Other)	3319	0.18	0.05	FAL: 689, TRU: 34
Do you usually add anything to your coffee? (No - just black)	82	0.98	0.66	TRU: 2611, FAL: 1349
Do you usually add anything to your coffee? (Milk, dairy alternative, or coffee creamer)	82	0.98	0.43	FAL: 2260, TRU: 1700
Do you usually add anything to your coffee? (Sugar or sweetener)	82	0.98	0.13	FAL: 3445, TRU: 515
Do you usually add anything to your coffee? (Flavor syrup)	82	0.98	0.06	FAL: 3729, TRU: 231
Do you usually add anything to your coffee? (Other)	82	0.98	0.01	FAL: 3914, TRU: 46
What kind of dairy do you add? (Whole milk)	2343	0.42	0.50	FAL: 852, TRU: 847
What kind of dairy do you add? (Skim milk)	2343	0.42	0.08	FAL: 1564, TRU: 135
What kind of dairy do you add? (Half and half)	2343	0.42	0.24	FAL: 1295, TRU: 404
What kind of dairy do you add? (Coffee creamer)	2343	0.42	0.09	FAL: 1550, TRU: 149
What kind of dairy do you add? (Flavored coffee creamer)	2343	0.42	0.10	FAL: 1537, TRU: 162
What kind of dairy do you add? (Oat milk)	2343	0.42	0.30	FAL: 1188, TRU: 511
What kind of dairy do you add? (Almond milk)	2343	0.42	0.09	FAL: 1554, TRU: 145
What kind of dairy do you add? (Soy milk)	2343	0.42	0.05	FAL: 1618, TRU: 81
What kind of dairy do you add? (Other)	2343	0.42	0.00	FAL: 1699
What kind of sugar or sweetener do you add? (Granulated Sugar)	3525	0.13	0.57	TRU: 293, FAL: 224
What kind of sugar or sweetener do you add? (Artificial Sweeteners (e.g., Splenda))	3525	0.13	0.18	FAL: 426, TRU: 91
What kind of sugar or sweetener do you add? (Honey)	3525	0.13	0.14	FAL: 447, TRU: 70
What kind of sugar or sweetener do you add? (Maple Syrup)	3525	0.13	0.07	FAL: 480, TRU: 37
What kind of sugar or sweetener do you add? (Stevia)	3525	0.13	0.10	FAL: 466, TRU: 51

skim_variable	n_missing	complete_rate	mean	count
What kind of sugar or sweetener do you add? (Agave Nectar)	3525	0.13	0.03	FAL: 502, TRU: 15
What kind of sugar or sweetener do you add? (Brown Sugar)	3525	0.13	0.14	FAL: 443, TRU: 74
What kind of sugar or sweetener do you add? (Raw Sugar (Turbinado))	3525	0.13	0.22	FAL: 401, TRU: 116
What kind of flavorings do you add?	4042	0.00	NaN	:
What kind of flavorings do you add? (Vanilla Syrup)	4042	0.00	NaN	:
What kind of flavorings do you add? (Caramel Syrup)	4042	0.00	NaN	:
What kind of flavorings do you add? (Hazelnut Syrup)	4042	0.00	NaN	:
What kind of flavorings do you add? (Cinnamon (Ground or Stick))	4042	0.00	NaN	:
What kind of flavorings do you add? (Peppermint Syrup)	4042	0.00	NaN	:
What kind of flavorings do you add? (Other)	4042	0.00	NaN	:
What other flavoring do you use?	4042	0.00	NaN	:
Why do you drink coffee? (It tastes good)	472	0.88	0.94	TRU: 3355, FAL: 215
Why do you drink coffee? (I need the caffeine)	472	0.88	0.57	TRU: 2021, FAL: 1549
Why do you drink coffee? (I need the ritual)	472	0.88	0.54	TRU: 1922, FAL: 1648
Why do you drink coffee? (It makes me go to the bathroom)	472	0.88	0.13	FAL: 3105, TRU: 465
Why do you drink coffee? (Other)	472	0.88	0.05	FAL: 3402, TRU: 168

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Lastly, how would you rate your own coffee expertise?	104	0.97	5.69	1.95	1	5	6	7	10	
Coffee A - Bitterness	244	0.94	2.14	0.95	1	1	2	3	5	
Coffee A - Acidity	263	0.93	3.63	0.98	1	3	4	4	5	
Coffee A - Personal Preference	253	0.94	3.31	1.19	1	2	3	4	5	
Coffee B - Bitterness	262	0.94	3.01	0.99	1	2	3	4	5	
Coffee B - Acidity	275	0.93	2.22	0.87	1	2	2	3	5	
Coffee B - Personal Preference	269	0.93	3.07	1.11	1	2	3	4	5	
Coffee C - Bitterness	278	0.93	3.07	1.00	1	2	3	4	5	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
Coffee C - Acidity	291	0.93	2.37	0.92	1	2	2	3	5	
Coffee C - Personal Preference	276	0.93	3.06	1.13	1	2	3	4	5	
Coffee D - Bitterness	275	0.93	2.16	1.08	1	1	2	3	5	
Coffee D - Acidity	277	0.93	3.86	1.01	1	3	4	5	5	
Coffee D - Personal Preference	278	0.93	3.38	1.45	1	2	4	5	5	

Data 2: Dating Experiences in Colleges

Introduction and data

- Identify the source of the data.

  The dataset was obtained from a speed dating experiment conducted at Columbia University, available at link: http://www.stat.columbia.edu/~gelman/arm/examples/speed.dating/

- State when and how it was originally collected (by the original data curator, not necessarily how you found the data).

  The data was originally collected through a series of speed dating events organized for Columbia University graduate and professional students. Participants were asked to fill out surveys before and after the speed dating sessions, as well as during follow-up periods.

- Write a brief description of the observations.

  The dataset includes information on participant demographics, preferences, and outcomes from the speed dating events. Variables include age, field of study, importance of attributes in a potential partner, and match outcomes.

Research question

- A well formulated research question. (You may include more than one research question if you want to receive feedback on different ideas for your project. However, one per data set is required.)
- A description of the research topic along with a concise statement of your hypotheses on this topic.
- Identify the types of variables in your research question. Categorical? Quantitative?

1. Main Question: What factors influence the likelihood of a match in speed dating among Columbia University students?

This study aims to investigate the factors that influence the likelihood of forming a match in speed dating events among Columbia University students. The focus is on understanding how individual preferences, demographic characteristics, and perceived attributes of potential partners contribute to successful matchmaking.

- Hypothesis 1: Participants are more likely to match with partners who share similar interests and hobbies.
- Hypothesis 2: The importance of physical attractiveness in a potential partner varies between genders, with males placing higher importance on this attribute compared to females.

Variables:

- Categorical: Gender, field of study, match outcome (yes/no)
- Quantitative: Age, importance ratings for various attributes (attractiveness, sincerity, intelligence, etc.), number of matches

2. Question: How do individual preferences and demographic characteristics influence the success rate of matches in speed dating events at Columbia University?

This research aims to explore the dynamics of speed dating at Columbia University, focusing on how participants' personal preferences and demographic factors affect their chances of forming successful matches.

- Hypothesis 1: Participants are more likely to form a match with individuals who share similar interests and values.
- Hypothesis 2: Age and field of study are significant predictors of match success rates, with participants preferring partners within similar age groups and academic disciplines.

Variables:

- Categorical Variables: Gender, field of study, match outcome (yes/no), race/ethnicity.
- Quantitative Variables: Age, ratings of importance for various attributes (e.g., attractiveness, intelligence), number of matches.

Glimpse of data

```
read_csv("data/Speed Dating Data.csv")|>
skim()
```

```
Rows: 8378 Columns: 195
-- Column specification --------------------------------------------------------
Delimiter: ","
chr   (4): field, undergra, from, career
dbl (187): iid, id, gender, idg, condtn, wave, round, position, positin1, or...
num   (4): mn_sat, tuition, zipcode, income

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.

Warning: There was 1 warning in `dplyr::summarize()`.
i In argument: `dplyr::across(tidyselect::any_of(variable_names),
  mangled_skimmers$funs)`.
i In group 0: .
Caused by warning:
! There were 27 warnings in `dplyr::summarize()`.
The first warning was:
i In argument: `dplyr::across(tidyselect::any_of(variable_names),
  mangled_skimmers$funs)`.
Caused by warning in `grepl()`:
! unable to translate 'Ecole Normale Sup<8e>rieure, Paris' to a wide string
i Run `dplyr::last_dplyr_warnings()` to see the 26 remaining warnings.
```

Table 5: Data summary

Name	read_csv(“data/Speed Dati…
Number of rows	8378
Number of columns	195
Column type frequency:	
character	4
numeric	191
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
field	63	0.99	3	56	0	259	0
undergra	3464	0.59	2	49	0	241	0

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
from	79	0.99	2	58	0	269	0
career	89	0.99	1	77	0	367	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
iid	0	1.00	283.68	158.58	1.00	154.00	281.00	407.00	552.00	
id	1	1.00	8.96	5.49	1.00	4.00	8.00	13.00	22.00	
gender	0	1.00	0.50	0.50	0.00	0.00	1.00	1.00	1.00	
idg	0	1.00	17.33	10.94	1.00	8.00	16.00	26.00	44.00	
condtn	0	1.00	1.83	0.38	1.00	2.00	2.00	2.00	2.00	
wave	0	1.00	11.35	6.00	1.00	7.00	11.00	15.00	21.00	
round	0	1.00	16.87	4.36	5.00	14.00	18.00	20.00	22.00	
position	0	1.00	9.04	5.51	1.00	4.00	8.00	13.00	22.00	
positin1	1846	0.78	9.30	5.65	1.00	4.00	9.00	14.00	22.00	
order	0	1.00	8.93	5.48	1.00	4.00	8.00	13.00	22.00	
partner	0	1.00	8.96	5.49	1.00	4.00	8.00	13.00	22.00	
pid	10	1.00	283.86	158.58	1.00	154.00	281.00	408.00	552.00	
match	0	1.00	0.16	0.37	0.00	0.00	0.00	0.00	1.00	
int_corr	158	0.98	0.20	0.30	-0.83	-0.02	0.21	0.43	0.91	
samerace	0	1.00	0.40	0.49	0.00	0.00	0.00	1.00	1.00	
age_o	104	0.99	26.36	3.56	18.00	24.00	26.00	28.00	55.00	
race_o	73	0.99	2.76	1.23	1.00	2.00	2.00	4.00	6.00	
pf_o_att	89	0.99	22.50	12.57	0.00	15.00	20.00	25.00	100.00	
pf_o_sin	89	0.99	17.40	7.04	0.00	15.00	18.37	20.00	60.00	
pf_o_int	89	0.99	20.27	6.78	0.00	17.39	20.00	23.81	50.00	
pf_o_fun	98	0.99	17.46	6.09	0.00	15.00	18.00	20.00	50.00	
pf_o_amb	107	0.99	10.69	6.13	0.00	5.00	10.00	15.00	53.00	
pf_o_sha	129	0.98	11.85	6.36	0.00	9.52	10.64	16.00	30.00	
dec_o	0	1.00	0.42	0.49	0.00	0.00	0.00	1.00	1.00	
attr_o	212	0.97	6.19	1.95	0.00	5.00	6.00	8.00	10.50	
sinc_o	287	0.97	7.18	1.74	0.00	6.00	7.00	8.00	10.00	
intel_o	306	0.96	7.37	1.55	0.00	6.00	7.00	8.00	10.00	
fun_o	360	0.96	6.40	1.95	0.00	5.00	7.00	8.00	11.00	
amb_o	722	0.91	6.78	1.79	0.00	6.00	7.00	8.00	10.00	
shar_o	1076	0.87	5.47	2.16	0.00	4.00	6.00	7.00	10.00	
like_o	250	0.97	6.13	1.84	0.00	5.00	6.00	7.00	10.00	
prob_o	318	0.96	5.21	2.13	0.00	4.00	5.00	7.00	10.00	
met_o	385	0.95	1.96	0.25	1.00	2.00	2.00	2.00	8.00	
age	95	0.99	26.36	3.57	18.00	24.00	26.00	28.00	55.00	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
field_cd	82	0.99	7.66	3.76	1.00	5.00	8.00	10.00	18.00	
mn_sat	5245	0.37	1299.66	119.80	914.00	1214.00	1310.00	1400.00	1490.00	
tuition	4795	0.43	21174.93	6748.66	2406.00	15162.00	25020.00	26562.00	34300.00	
race	63	0.99	2.76	1.23	1.00	2.00	2.00	4.00	6.00	
imprace	79	0.99	3.78	2.85	0.00	1.00	3.00	6.00	10.00	
imprelig	79	0.99	3.65	2.81	1.00	1.00	3.00	6.00	10.00	
zipcode	1064	0.87	75423.22	525492.44	0.00	10021.00	19041.00	75840.75	9971200.00	
income	4099	0.51	44887.61	17206.92	8607.00	31516.00	43185.00	54303.00	109031.00	
goal	79	0.99	2.12	1.41	1.00	1.00	2.00	2.00	6.00	
date	97	0.99	5.01	1.44	1.00	4.00	5.00	6.00	7.00	
go_out	79	0.99	2.16	1.11	1.00	1.00	2.00	3.00	7.00	
career_c	138	0.98	5.28	3.31	1.00	2.00	6.00	7.00	17.00	
sports	79	0.99	6.43	2.62	1.00	4.00	7.00	9.00	10.00	
tvsports	79	0.99	4.58	2.80	1.00	2.00	4.00	7.00	10.00	
exercise	79	0.99	6.25	2.42	1.00	5.00	6.00	8.00	10.00	
dining	79	0.99	7.78	1.75	1.00	7.00	8.00	9.00	10.00	
museums	79	0.99	6.99	2.05	0.00	6.00	7.00	9.00	10.00	
art	79	0.99	6.71	2.26	0.00	5.00	7.00	8.00	10.00	
hiking	79	0.99	5.74	2.57	0.00	4.00	6.00	8.00	10.00	
gaming	79	0.99	3.88	2.62	0.00	2.00	3.00	6.00	14.00	
clubbing	79	0.99	5.75	2.50	0.00	4.00	6.00	8.00	10.00	
reading	79	0.99	7.68	2.01	1.00	7.00	8.00	9.00	13.00	
tv	79	0.99	5.30	2.53	1.00	3.00	6.00	7.00	10.00	
theater	79	0.99	6.78	2.24	0.00	5.00	7.00	9.00	10.00	
movies	79	0.99	7.92	1.70	0.00	7.00	8.00	9.00	10.00	
concerts	79	0.99	6.83	2.16	0.00	5.00	7.00	8.00	10.00	
music	79	0.99	7.85	1.79	1.00	7.00	8.00	9.00	10.00	
shopping	79	0.99	5.63	2.61	1.00	4.00	6.00	8.00	10.00	
yoga	79	0.99	4.34	2.72	0.00	2.00	4.00	7.00	10.00	
exphappy	101	0.99	5.53	1.73	1.00	5.00	6.00	7.00	10.00	
expnum	6578	0.21	5.57	4.76	0.00	2.00	4.00	8.00	20.00	
attr1_1	79	0.99	22.51	12.59	0.00	15.00	20.00	25.00	100.00	
sinc1_1	79	0.99	17.40	7.05	0.00	15.00	18.18	20.00	60.00	
intel1_1	79	0.99	20.27	6.78	0.00	17.39	20.00	23.81	50.00	
fun1_1	89	0.99	17.46	6.09	0.00	15.00	18.00	20.00	50.00	
amb1_1	99	0.99	10.68	6.12	0.00	5.00	10.00	15.00	53.00	
shar1_1	121	0.99	11.85	6.36	0.00	9.52	10.64	16.00	30.00	
attr4_1	1889	0.77	26.39	16.30	5.00	10.00	25.00	35.00	95.00	
sinc4_1	1889	0.77	11.07	6.66	0.00	6.00	10.00	15.00	35.00	
intel4_1	1889	0.77	12.64	6.72	0.00	8.00	10.00	16.00	35.00	
fun4_1	1889	0.77	15.57	7.33	0.00	10.00	15.00	20.00	45.00	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
amb4_1	1889	0.77	9.78	7.00	0.00	5.00	10.00	15.00	50.00	
shar4_1	1911	0.77	11.01	6.06	0.00	7.00	10.00	15.00	40.00	
attr2_1	79	0.99	30.36	16.25	0.00	20.00	25.00	40.00	100.00	
sinc2_1	79	0.99	13.27	6.98	0.00	10.00	15.00	18.75	50.00	
intel2_1	79	0.99	14.42	6.26	0.00	10.00	15.00	20.00	40.00	
fun2_1	79	0.99	18.42	6.58	0.00	15.00	20.00	20.00	50.00	
amb2_1	89	0.99	11.74	6.89	0.00	6.00	10.00	15.00	50.00	
shar2_1	89	0.99	11.85	6.17	0.00	10.00	10.00	15.63	30.00	
attr3_1	105	0.99	7.08	1.40	2.00	6.00	7.00	8.00	10.00	
sinc3_1	105	0.99	8.29	1.41	2.00	8.00	8.00	9.00	10.00	
fun3_1	105	0.99	7.70	1.56	2.00	7.00	8.00	9.00	10.00	
intel3_1	105	0.99	8.40	1.08	3.00	8.00	8.00	9.00	10.00	
amb3_1	105	0.99	7.58	1.78	2.00	7.00	8.00	9.00	10.00	
attr5_1	3472	0.59	6.94	1.50	2.00	6.00	7.00	8.00	10.00	
sinc5_1	3472	0.59	7.93	1.63	1.00	7.00	8.00	9.00	10.00	
intel5_1	3472	0.59	8.28	1.28	3.00	8.00	8.00	9.00	10.00	
fun5_1	3472	0.59	7.43	1.78	2.00	6.00	8.00	9.00	10.00	
amb5_1	3472	0.59	7.62	1.77	1.00	7.00	8.00	9.00	10.00	
dec	0	1.00	0.42	0.49	0.00	0.00	0.00	1.00	1.00	
attr	202	0.98	6.19	1.95	0.00	5.00	6.00	8.00	10.00	
sinc	277	0.97	7.18	1.74	0.00	6.00	7.00	8.00	10.00	
intel	296	0.96	7.37	1.55	0.00	6.00	7.00	8.00	10.00	
fun	350	0.96	6.40	1.95	0.00	5.00	7.00	8.00	10.00	
amb	712	0.92	6.78	1.79	0.00	6.00	7.00	8.00	10.00	
shar	1067	0.87	5.47	2.16	0.00	4.00	6.00	7.00	10.00	
like	240	0.97	6.13	1.84	0.00	5.00	6.00	7.00	10.00	
prob	309	0.96	5.21	2.13	0.00	4.00	5.00	7.00	10.00	
met	375	0.96	0.95	0.99	0.00	0.00	0.00	2.00	8.00	
match_es	1173	0.86	3.21	2.44	0.00	2.00	3.00	4.00	18.00	
attr1_s	4282	0.49	20.79	12.97	3.00	14.81	17.65	25.00	95.00	
sinc1_s	4282	0.49	15.43	6.92	0.00	10.00	15.79	20.00	50.00	
intel1_s	4282	0.49	17.24	6.60	0.00	10.00	18.42	20.00	40.00	
fun1_s	4282	0.49	15.26	5.36	1.00	10.00	15.91	20.00	40.00	
amb1_s	4282	0.49	11.14	5.51	0.00	7.00	10.00	15.00	23.81	
shar1_s	4282	0.49	12.46	5.92	0.00	9.00	12.50	16.28	30.00	
attr3_s	4378	0.48	7.21	1.42	3.00	7.00	7.00	8.00	10.00	
sinc3_s	4378	0.48	8.08	1.46	1.00	7.00	8.00	9.00	10.00	
intel3_s	4378	0.48	8.26	1.18	4.00	8.00	8.00	9.00	10.00	
fun3_s	4378	0.48	7.69	1.63	3.00	7.00	8.00	9.00	10.00	
amb3_s	4378	0.48	7.59	1.79	2.00	7.00	8.00	9.00	10.00	
satis_2	915	0.89	5.71	1.82	1.00	5.00	6.00	7.00	10.00	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
length	915	0.89	1.84	0.98	1.00	1.00	1.00	3.00	3.00	
numdat_2	945	0.89	2.34	0.63	1.00	2.00	2.00	3.00	3.00	
attr7_2	6394	0.24	32.82	17.16	10.00	20.00	30.00	40.00	80.00	
sinc7_2	6423	0.23	13.53	7.98	0.00	10.00	10.00	20.00	40.00	
intel7_2	6394	0.24	15.29	7.29	0.00	10.00	15.00	20.00	50.00	
fun7_2	6394	0.24	18.87	8.54	0.00	10.00	20.00	24.00	50.00	
amb7_2	6423	0.23	7.29	6.13	0.00	0.00	5.00	10.00	20.00	
shar7_2	6404	0.24	12.16	8.24	0.00	5.00	10.00	20.00	40.00	
attr1_2	933	0.89	26.22	14.39	5.00	16.67	20.00	30.00	85.00	
sinc1_2	915	0.89	15.87	6.66	0.00	10.00	16.67	20.00	50.00	
intel1_2	915	0.89	17.81	6.54	0.00	15.00	19.05	20.00	40.00	
fun1_2	915	0.89	17.65	6.13	0.00	15.00	18.37	20.00	50.00	
amb1_2	915	0.89	9.91	5.68	0.00	5.00	10.00	15.00	22.22	
shar1_2	915	0.89	12.76	6.65	0.00	10.00	13.00	16.67	35.00	
attr4_2	2603	0.69	26.81	16.40	6.00	10.00	25.00	40.00	100.00	
sinc4_2	2603	0.69	11.93	6.40	0.00	8.00	10.00	15.00	35.00	
intel4_2	2603	0.69	12.10	5.99	0.00	8.00	10.00	15.00	40.00	
fun4_2	2603	0.69	15.16	7.29	0.00	9.00	15.00	20.00	50.00	
amb4_2	2603	0.69	9.34	5.86	0.00	5.00	10.00	10.00	35.00	
shar4_2	2603	0.69	11.32	6.30	0.00	7.00	10.00	15.00	40.00	
attr2_2	2603	0.69	29.34	14.55	0.00	19.15	25.00	38.46	85.00	
sinc2_2	2603	0.69	13.90	6.17	0.00	10.00	15.00	19.23	40.00	
intel2_2	2603	0.69	13.96	5.40	0.00	10.00	15.00	17.39	30.77	
fun2_2	2603	0.69	17.97	6.10	0.00	15.00	18.52	20.00	40.00	
amb2_2	2603	0.69	11.91	6.31	0.00	10.00	10.00	15.09	50.00	
shar2_2	2603	0.69	12.89	5.62	0.00	10.00	13.95	16.52	30.00	
attr3_2	915	0.89	7.13	1.37	2.00	7.00	7.00	8.00	10.00	
sinc3_2	915	0.89	7.93	1.50	2.00	7.00	8.00	9.00	10.00	
intel3_2	915	0.89	8.24	1.18	4.00	8.00	8.00	9.00	10.00	
fun3_2	915	0.89	7.60	1.55	1.00	7.00	8.00	9.00	10.00	
amb3_2	915	0.89	7.49	1.74	2.00	7.00	8.00	9.00	10.00	
attr5_2	4001	0.52	6.83	1.41	2.00	6.00	7.00	8.00	10.00	
sinc5_2	4001	0.52	7.39	1.59	2.00	6.00	8.00	8.00	10.00	
intel5_2	4001	0.52	7.84	1.28	2.00	7.00	8.00	9.00	10.00	
fun5_2	4001	0.52	7.28	1.65	2.00	6.00	7.00	8.00	10.00	
amb5_2	4001	0.52	7.33	1.52	2.00	6.00	7.00	8.00	10.00	
you_call	4404	0.47	0.78	1.61	0.00	0.00	0.00	1.00	21.00	
them_cal	4404	0.47	0.98	1.38	0.00	0.00	1.00	1.00	9.00	
date_3	4404	0.47	0.38	0.48	0.00	0.00	0.00	1.00	1.00	
numdat_3	6882	0.18	1.23	1.29	0.00	1.00	1.00	1.00	9.00	
num_in_3	7710	0.08	0.93	0.75	0.00	1.00	1.00	1.00	4.00	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
attr1_3	4404	0.47	24.38	13.71	0.00	15.22	20.00	30.00	80.00	
sinc1_3	4404	0.47	16.59	7.47	0.00	10.00	16.67	20.00	65.00	
intel1_3	4404	0.47	19.41	6.12	0.00	16.67	20.00	20.00	45.00	
fun1_3	4404	0.47	16.23	5.16	0.00	14.81	16.33	20.00	30.00	
amb1_3	4404	0.47	10.90	5.90	0.00	5.00	10.00	15.00	30.00	
shar1_3	4404	0.47	12.70	6.56	0.00	10.00	14.29	16.67	55.00	
attr7_3	6362	0.24	31.33	17.55	0.00	20.00	25.00	40.00	80.00	
sinc7_3	6362	0.24	15.65	9.34	0.00	10.00	15.00	20.00	60.00	
intel7_3	6362	0.24	16.68	7.88	0.00	10.00	18.00	20.00	45.00	
fun7_3	6362	0.24	16.42	7.23	0.00	10.00	17.00	20.00	40.00	
amb7_3	6362	0.24	7.82	6.10	0.00	0.00	10.00	10.00	30.00	
shar7_3	6362	0.24	12.21	8.62	0.00	5.00	10.00	20.00	55.00	
attr4_3	5419	0.35	25.61	17.48	0.00	10.00	20.00	37.00	80.00	
sinc4_3	5419	0.35	10.75	5.74	0.00	7.00	10.00	15.00	40.00	
intel4_3	5419	0.35	11.52	6.00	0.00	7.00	10.00	15.00	30.00	
fun4_3	5419	0.35	14.28	6.93	0.00	9.00	12.00	20.00	30.00	
amb4_3	5419	0.35	9.21	6.39	0.00	5.00	9.00	10.00	40.00	
shar4_3	5419	0.35	11.25	6.52	0.00	7.00	10.00	15.00	45.00	
attr2_3	5419	0.35	24.97	17.01	5.00	10.00	20.00	35.00	80.00	
sinc2_3	5419	0.35	10.92	6.23	0.00	7.00	10.00	15.00	50.00	
intel2_3	5419	0.35	11.95	7.01	0.00	7.00	10.00	15.00	60.00	
fun2_3	5419	0.35	14.96	7.94	0.00	9.00	15.00	20.00	40.00	
amb2_3	5419	0.35	9.53	6.40	0.00	6.00	10.00	10.00	50.00	
shar2_3	6362	0.24	11.97	7.01	0.00	5.00	10.00	15.00	45.00	
attr3_3	4404	0.47	7.24	1.58	2.00	7.00	7.00	8.00	12.00	
sinc3_3	4404	0.47	8.09	1.61	2.00	7.00	8.00	9.00	12.00	
intel3_3	4404	0.47	8.39	1.46	3.00	8.00	8.00	9.00	12.00	
fun3_3	4404	0.47	7.66	1.74	2.00	7.00	8.00	9.00	12.00	
amb3_3	4404	0.47	7.39	1.96	1.00	6.00	8.00	9.00	12.00	
attr5_3	6362	0.24	6.81	1.51	2.00	6.00	7.00	8.00	10.00	
sinc5_3	6362	0.24	7.62	1.50	2.00	7.00	8.00	9.00	10.00	
intel5_3	6362	0.24	7.93	1.34	4.00	7.00	8.00	9.00	10.00	
fun5_3	6362	0.24	7.16	1.67	1.00	6.00	7.00	8.00	10.00	
amb5_3	6362	0.24	7.05	1.72	1.00	6.00	7.00	8.00	10.00	

Data 3

Introduction and data

- Identify the source of the data.

  The dataset was obtained from IMDb (Internet Movie Database), specifically from their non-commercial datasets page. The data is available for download at https://datasets.imdbws.com/.

- State when and how it was originally collected (by the original data curator, not necessarily how you found the data).

  IMDb collects data from a variety of sources, including studios, filmmakers, and viewers. The dataset is updated daily and includes information from IMDb's extensive database of movies, TV shows, and other entertainment content.

- Write a brief description of the observations.

  The dataset includes several files, each containing different types of information such as titles, crew, episodes, ratings, and principal cast members. Each file is in a gzipped, tab-separated-values (TSV) format and includes headers that describe the contents of each column.

Research question

- A well formulated research question. (You may include more than one research question if you want to receive feedback on different ideas for your project. However, one per data set is required.)
- A description of the research topic along with a concise statement of your hypotheses on this topic.
- Identify the types of variables in your research question. Categorical? Quantitative?

1. Research Question: How does film and genre views and popularity differ across different video platforms? How does trends in film popularity across different platforms correlate with differing demographic characteristics of platform users?

This study aims to investigate the trends of film ratings/genre popularity, and the number of views across different streaming platforms, as indicated by IMDb ratings and viewer preferences.

The focus is on understanding which platforms are preferred by viewers for different types of content and how these preferences are reflected in content ratings.

- Hypothesis 1: Netflix is the most popular streaming platform among IMDb users, as indicated by the highest average ratings for its exclusive content.
- Hypothesis 2: The popularity of streaming platforms varies by genre, with certain platforms being preferred for specific types of content. Different audience groups that a platform is tailored to also influence media popularity rankings within a platform.

Variables:

- Categorical: Title, isAdult (whether it is an adult film), genre
- Quantitative: Average rating, number of votes, movie runtime, start and end year

2. Research Question: How does movie ratings across different platforms change over time?

- This study will investigate the trends of movie views across different streaming patterns over time, the processes through which different movies gain or lose popularity, and how patterns of trending content differ across different platforms

Hypothesis: Trends of movie popularity will be approximately comparable across different video platforms through time.

- Categorical: Title, isAdult (whether it is an adult film), genre
- Quantitative: Average rating, number of votes, movie runtime, start and end year

Glimpse of data

```
#imdb_basics <- read_tsv("data/title.basics.tsv.tsv")
#imdb_ratings <- read_tsv("data/title.ratings.tsv.tsv")
#imdb_data <- inner_join(
#  x = imdb_basics, y = imdb_ratings,
#)
#glimpse(imdb_data)

#Note: we had trouble uploading the imdb_data dataset due to an error: file size is 869.3 MB

#1,412,614
#Columns: 11
#$ tconst         <chr> "tt0000001", "tt0000002", "tt000…
#$ titleType      <chr> "short", "short", "short", "shor…
#$ primaryTitle   <chr> "Carmencita", "Le clown et ses c…
#$ originalTitle  <chr> "Carmencita", "Le clown et ses c…
#$ isAdult        <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,…
#$ startYear      <chr> "1894", "1892", "1892", "1892", …
#$ endYear        <chr> "\\N", "\\N", "\\N", "\\N", "\\N…
#$ runtimeMinutes <chr> "1", "5", "4", "12", "1", "1", "…
#$ genres         <chr> "Documentary,Short", "Animation,…
#$ averageRating  <dbl> 5.7, 5.7, 6.5, 5.4, 6.2, 5.0, 5.…
#$ numVotes       <dbl> 2032, 272, 1977, 178, 2735, 183,…
```