Project title

Proposal

Brilliant Cassowary

```
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 a coffee industry research group. 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

```
# add code here
```

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

```
# add code here
```

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?

  Main 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? How does movie ratings acoss different platforms change over time?

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, 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,…
```