DATA 607 - Assignment 2

Shoshana Farber

February 5, 2023

Connecting to PostgreSQL Database

```
my_pass <- read_file("C:/Users/Shoshana/Documents/pass.txt")

con <- dbConnect(
  Postgres(),
  host = "localhost",
  port = 5432,
  user = "postgres",
  password = my_pass,
  dbname = "cuny-sps"
)
```

Loading the Databases

```
movie_ratings <- dbGetQuery(con, "SELECT * FROM movie_ratings")
raters <- dbGetQuery(con, "SELECT * FROM raters")
movies <- dbGetQuery(con, "SELECT * FROM movies")

# preview each table
head(movie_ratings)
```

```
##   raterid movieid rating
## 1       1       1     NA
## 2       1       2      3
## 3       1       3     NA
## 4       1       4      3
## 5       1       5     NA
## 6       1       6     NA
```

```
head(raters)
```

```
##   raterid   name age
## 1       1  Sarah  24
## 2       2  Shani  22
## 3       3   Leah  61
## 4       4 Shimon  34
## 5       5  Dinah  23
## 6       6    Abe  61
```

```
head(movies)
```

```
##   movieid                   movie_title release_date
## 1       1       Avatar The Way of Water   2022-12-16
## 2       2 Black Panther Wakanda Forever   2022-11-11
## 3       3        Knives Out Glass Onion   2022-09-10
## 4       4           Matilda the Musical   2022-12-02
## 5       5              Top Gun Maverick   2022-05-27
## 6       6                  Bullet Train   2022-08-05
```

```
# want a table of just the raters, movies, and their ratings
movie_ratings <- movie_ratings %>%
  left_join(movies, on = "movieID") %>%
  left_join(raters, on = "raterID") %>%
  transmute(name, movie_title, rating)
```

```
## Joining, by = "movieid"
## Joining, by = "raterid"
```

```
head(movie_ratings)
```

```
##    name                   movie_title rating
## 1 Sarah       Avatar The Way of Water     NA
## 2 Sarah Black Panther Wakanda Forever      3
## 3 Sarah        Knives Out Glass Onion     NA
## 4 Sarah           Matilda the Musical      3
## 5 Sarah              Top Gun Maverick     NA
## 6 Sarah                  Bullet Train     NA
```

Exploration

What is the average rating for each movie?

```
avg_rating <- movie_ratings %>%
  group_by(movie_title) %>%
  filter(rating != is.na(rating)) %>%
  summarize(avg_rating = round(mean(rating), 2))

avg_rating
```

```
## # A tibble: 8 x 2
##   movie_title                   avg_rating
##   <chr>                              <dbl>
## 1 Avatar The Way of Water             4.17
## 2 Black Panther Wakanda Forever       3
## 3 Bullet Train                        3.5
## 4 Don't Worry Darling                 4
## 5 Knives Out Glass Onion              3
## 6 Matilda the Musical                 4
## 7 Ticket to Paradise                  4.5
## 8 Top Gun Maverick                    3.71
```

Which movie is highest rated?

```
avg_rating %>%
  arrange(-avg_rating)
```

```
## # A tibble: 8 x 2
##   movie_title                   avg_rating
##   <chr>                              <dbl>
## 1 Ticket to Paradise                  4.5
## 2 Avatar The Way of Water             4.17
## 3 Don't Worry Darling                 4
## 4 Matilda the Musical                 4
## 5 Top Gun Maverick                    3.71
## 6 Bullet Train                        3.5
## 7 Black Panther Wakanda Forever       3
## 8 Knives Out Glass Onion              3
```

Ticket to Paradise is the highest rated movie.

How many people watched the highest rated movie?

```
movie_ratings %>%
  filter(movie_title == "Ticket to Paradise",
         rating != is.na(rating))
```

```
##    name        movie_title rating
## 1 Dinah Ticket to Paradise      5
## 2 Leeor Ticket to Paradise      4
```

So ticket to paradise is the highest rated, but only 2/8 people actually watched it.

Which movie did most people watch?

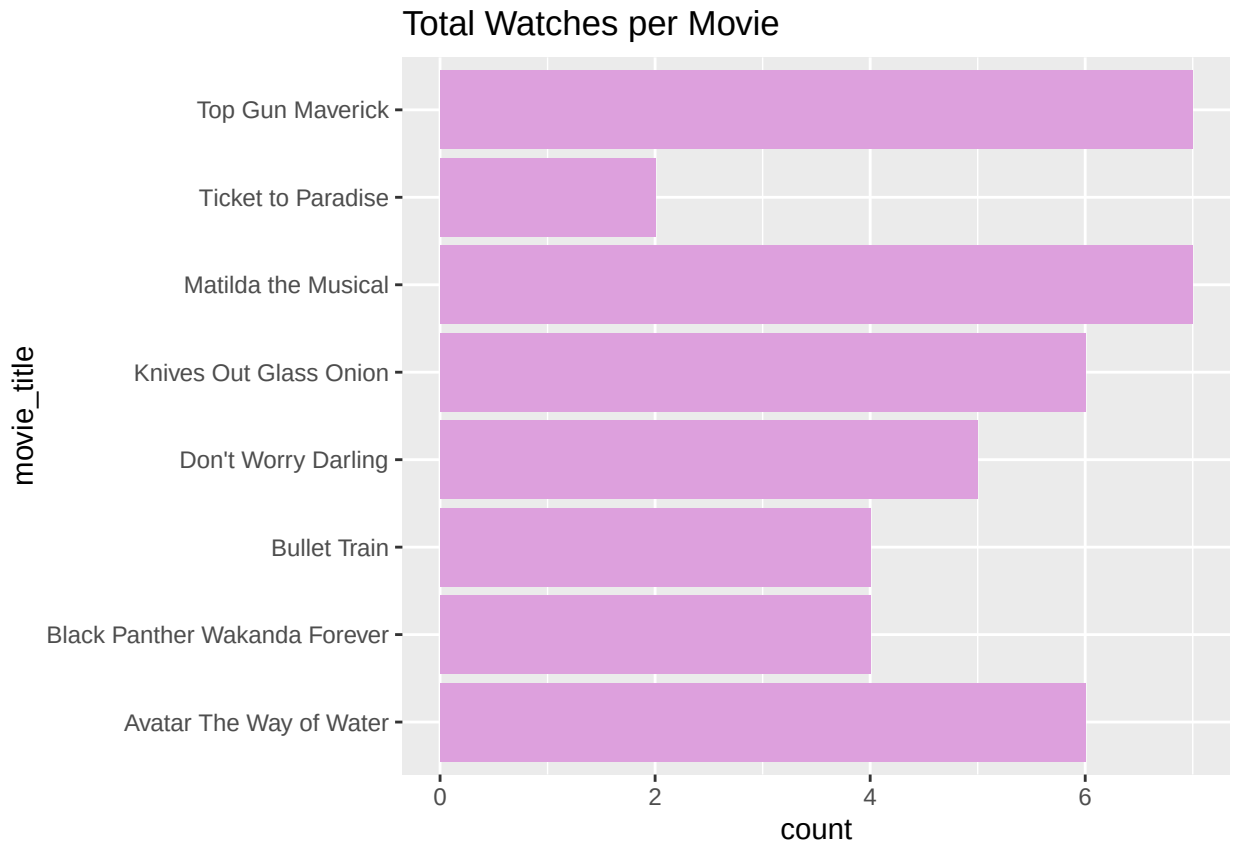
```
watched <- movie_ratings %>%
  filter(rating != is.na(rating)) %>%
  mutate(watched = 1)

watched %>%
  group_by(movie_title) %>%
  summarize(num_watched = sum(watched)) %>%
  arrange(-num_watched)
```

```
## # A tibble: 8 x 2
##   movie_title                   num_watched
##   <chr>                               <dbl>
## 1 Matilda the Musical                     7
## 2 Top Gun Maverick                        7
```

```
## 3 Avatar The Way of Water                6
## 4 Knives Out Glass Onion                 6
## 5 Don't Worry Darling                    5
## 6 Black Panther Wakanda Forever          4
## 7 Bullet Train                           4
## 8 Ticket to Paradise                     2
```

```
watched %>%
  ggplot(aes(y = movie_title)) +
    geom_bar(fill = "plum") +
    labs(title = "Total Watches per Movie")
```



Matilda the Musical and Top Gun Maverick were watched by the most people (7/8).

What is the average rating given by each person?

```
avg_rating_person <- movie_ratings %>%
  group_by(name) %>%
  filter(rating != is.na(rating)) %>%
  summarize(avg_rating = round(mean(rating), 2))

avg_rating_person
```

```
## # A tibble: 8 x 2
```

```
##   name   avg_rating
##   <chr>       <dbl>
## 1 Abe          3.6
## 2 Dinah        4
## 3 Leah         4
## 4 Leeor        4.17
## 5 Sarah        2.33
## 6 Shani        4
## 7 Shimon       3.4
## 8 Talia        3.5
```

Explanation

I connected to PostgreSQL server to access the `movie_ratings`, `movies`, and `raters` schemas directly from the database. `raters` has a primary key of **raterID**, and `movies` has a primary key of **movieID**. These are both foreign keys in `movie_ratings`.

To analyze the data here, I reassigned `movie_ratings` to be a table of just the names of the raters, the movie titles, and their respective ratings. If I had more data, I would analyze popularity based on age, as the `raters` table includes the ages of those who rated. Additionally, all these movies are recent releases and were released in the same year, but if I had many more movies from many more years I would want to group based on year to see most popular movies per year.