

Midterm 1 Batch B

Questions

Table of contents

Penguins	1
Question 1	2
NYC Flights	2
Question 2	4
Question 3	4
Countries and populations	5
Question 4	5
Question 5	5
Duke Forest houses	6
Question 6	7
Question 7	7
Law & Order	7
Question 8	8
Question 9	9
Romance and comedy	9
Question 10	10
Data	10
Question 11	11
Question 12	13
Question 13	13
Question 14	14

Penguins

The `penguins` data set includes measurements for penguin species, including: flipper length, body mass, bill dimensions, and sex. The following table summarizes information on which species of penguins (Adelie, Gentoo, and Chinstrap) live on which islands (Biscoe, Dream, or Torgersen).

Island	Adelie	Gentoo	Chinstrap	Total
Biscoe	44	124	0	168
Dream	56	0	68	124
Torgersen	52	0	0	52
Total	152	124	68	344

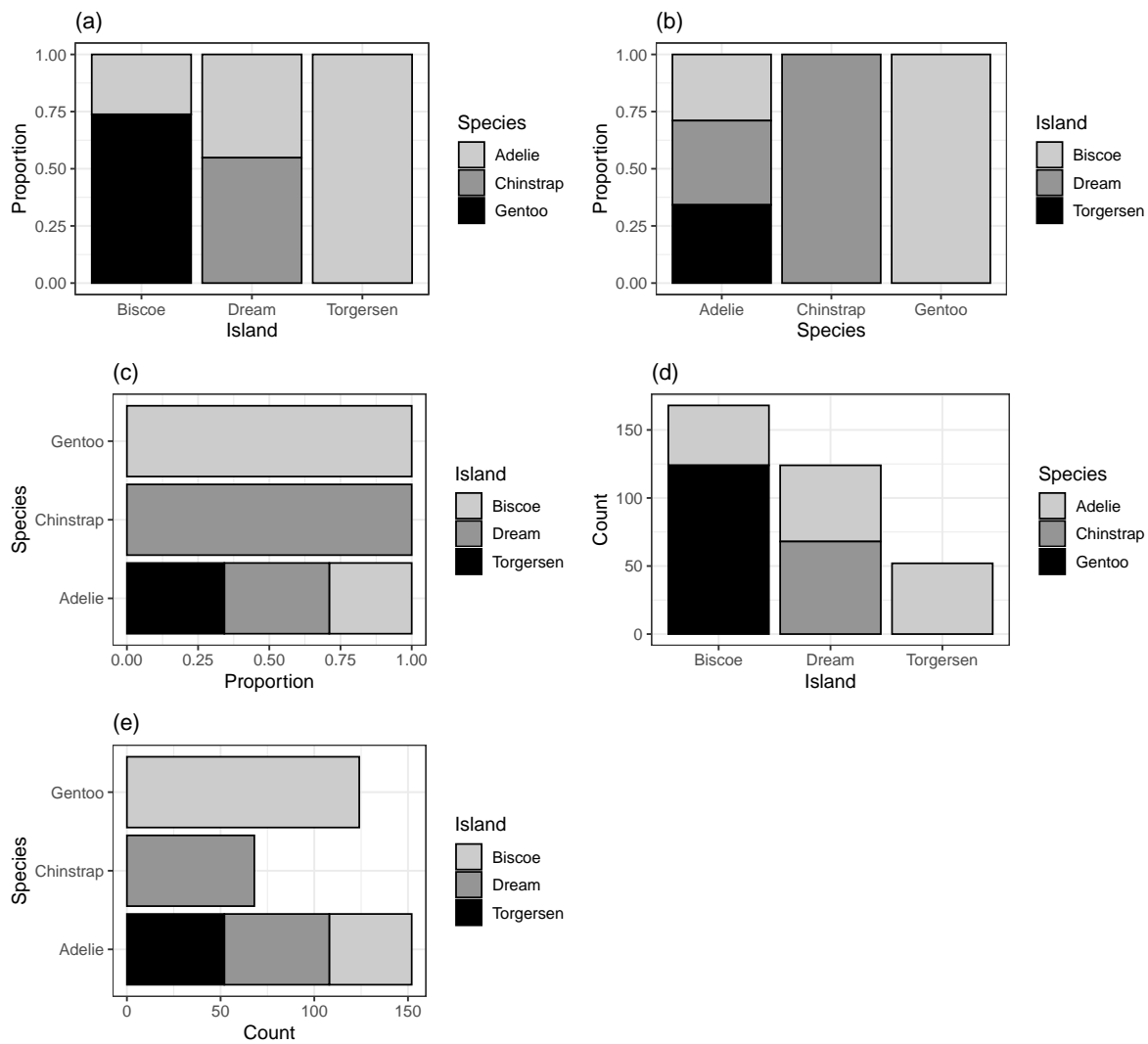
Question 1

Which of the following plots is the result of the following code?

NYC Flights

The `flights` dataset includes characteristics of all flights departing from New York City airports (JFK, LGA, EWR) in 2013. Below is a peek at the first ten rows of the `flights` data.

```
# A tibble: 336,776 x 19
  year month   day arr_delay carrier dep_time sched_dep_time dep_delay
  <int> <int> <int>    <dbl> <chr>      <int>         <int>      <dbl>
1  2013     1     1        11 UA           517           515         2
2  2013     1     1        20 UA           533           529         4
3  2013     1     1        33 AA           542           540         2
4  2013     1     1       -18 B6           544           545        -1
5  2013     1     1       -25 DL           554           600        -6
6  2013     1     1        12 UA           554           558        -4
7  2013     1     1        19 B6           555           600        -5
8  2013     1     1       -14 EV           557           600        -3
9  2013     1     1        -8 B6           557           600        -3
10 2013     1     1         8 AA           558           600        -2
# i 336,766 more rows
# i 11 more variables: arr_time <int>, sched_arr_time <int>, flight <int>,
#   tailnum <chr>, origin <chr>, dest <chr>, air_time <dbl>, distance <dbl>,
#   hour <dbl>, minute <dbl>, time_hour <dtm>
```



Question 2

Based on this output, which of the following must be true about the `flights` data frame? **Select all that are true.**

- a. The `flights` data frame is a `tibble`.
- b. The `flights` data frame has 10 rows.
- c. The `flights` data frame has 8 columns.
- d. The `carrier` variable in the `flights` data frame is a character variable.
- e. There are no missing data in the `flights` data frame.

Question 3

Which of the following pipelines produce(s) the output shown below? **Select all that apply.**

```
# A tibble: 336,776 x 5
  arr_delay carrier  year month  day
    <dbl>   <chr>    <int> <int> <int>
1     1272 HA      2013     1     9
2     1127 MQ      2013     6    15
3     1109 MQ      2013     1    10
4     1007 AA      2013     9    20
5      989 MQ      2013     7    22
6      931 DL      2013     4    10
7      915 DL      2013     3    17
8      895 DL      2013     7    22
9      878 AA      2013    12     5
10     875 MQ      2013     5     3
# i 336,766 more rows
```

- a.
- b.
- c.
- d.
- e.

Countries and populations

We have a small dataset of six countries and their populations:

```
# A tibble: 6 x 2
  country      population
  <chr>         <dbl>
1 Curacao         150
2 Ecuador       18001
3 Iraq          44496.
4 New Zealand     5124.
5 Palau           18.0
6 United States 333288.
```

And another small dataset of five countries and the continent they're in:

```
# A tibble: 5 x 3
  entity      code continent
  <chr>      <chr> <chr>
1 Angola     AGO   Africa
2 Curacao    CUW   North America
3 Ecuador    ECU   South America
4 Iraq       IRQ   Asia
5 New Zealand NZL   Oceania
```

You join the two datasets with the following:

Question 4

How many rows will the resulting data frame have?

- a. 4 b. 5 c. 6 d. 7 e. 8

Question 5

What will be the columns of the resulting data frame?

- a. country, population
b. country, population, code, continent
c. entity, code, continent

	Built earlier than 1950	Built in 1950 or later
Garage	5	33
No garage	3	57

d. entity, population, code, continent

e. country, entity, population, code, continent

Duke Forest houses

The `duke_forest` dataset includes information on prices and various other features (number of bedrooms, bathrooms, area, year built, type of cooling, type of heating, etc.) of houses in the Duke Forest neighborhood of Durham, NC.

Rows: 98

Columns: 13

```
$ address    <chr> "1 Learned Pl, Durham, NC 27705", "1616 Pinecrest Rd, Durha~
$ price      <dbl> 1520000, 1030000, 420000, 680000, 428500, 456000, 1270000, ~
$ bed        <dbl> 3, 5, 2, 4, 4, 3, 5, 4, 4, 3, 4, 4, 3, 5, 4, 5, 3, 4, 4, 3,~
$ bath       <dbl> 4.0, 4.0, 3.0, 3.0, 3.0, 3.0, 5.0, 3.0, 5.0, 2.0, 3.0, 3.0,~
$ area       <dbl> 6040, 4475, 1745, 2091, 1772, 1950, 3909, 2841, 3924, 2173,~
$ type       <chr> "Single Family", "Single Family", "Single Family", "Single ~
$ year_built <dbl> 1972, 1969, 1959, 1961, 2020, 2014, 1968, 1973, 1972, 1964,~
$ heating    <chr> "Other, Gas", "Forced air, Gas", "Forced air, Gas", "Heat p~
$ cooling     <fct> central, central, central, central, central, central, centr~
$ parking    <chr> "0 spaces", "Carport, Covered", "Garage - Attached, Covered~
$ lot        <dbl> 0.97, 1.38, 0.51, 0.84, 0.16, 0.45, 0.94, 0.79, 0.53, 0.73,~
$ hoa        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA,~
$ url        <chr> "https://www.zillow.com/homedetails/1-Learned-Pl-Durham-NC--"
```

The following summary table gives us some information about whether homes in this data set have garages and when they were built.

See next page for questions on this dataset.

The pipeline below produces a data frame with a fewer number of rows than `duke_forest`.

```
# A tibble: 5 x 5
  parking year_built price area price_per_sqfeet
  <chr>      <dbl>  <dbl> <dbl>      <dbl>
1 Garage    1945 900000  2933      307.
2 Garage    1938 265000  1300      204.
3 Garage    1934 600000  2514      239.
4 Garage    1941 412500  1661      248.
5 Garage    1940 105000  1094       96.0
```

Question 6

Which of the following goes in blanks (1) and (2)?

	(1)	(2)
a.	&	<
b.		<
c.	&	>=
d.		>=
e.	&	!=

Question 7

Which function or functions go into blank (3)? **Select all that apply.**

- a. `arrange()`
- b. `mutate()`
- c. `filter()`
- d. `summarize()`
- e. `slice()`

Law & Order

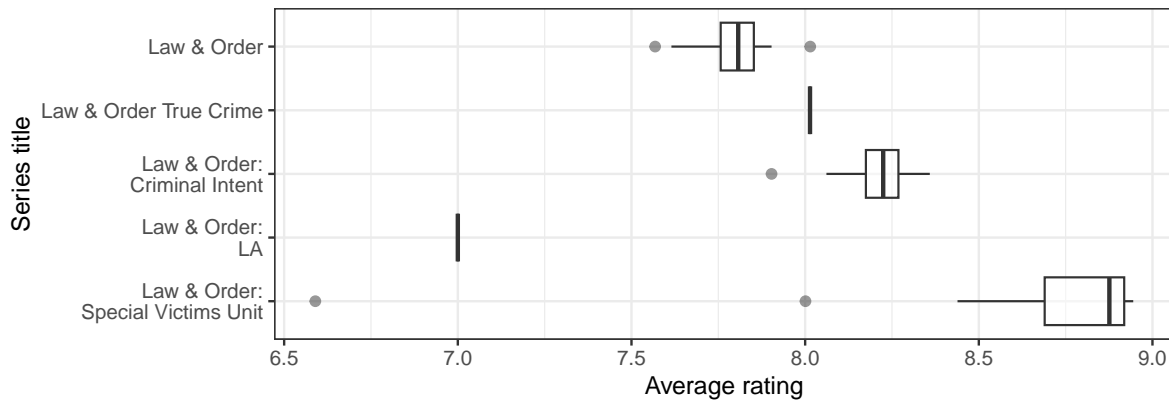
You've heard of the tidyverse, now let's visit the Law & Order-verse. Doink doink!¹

¹“Doink doink” is the scene and episode introductory sound on the Law & Order series. If you've never heard it, you're not at any disadvantage for the exam. If you've ever heard it, good luck getting it out of your head!

Law & Order is a police procedural and legal drama television series that has been running since the 1990s. The Law & Order franchise includes a number of series such as Law & Order, Law & Order: SVU, Law & Order: Criminal Intent, etc.

You will work with data on average ratings for each season of three series from the Law & Order-verse – a subset of the data from the previous questions. Below is a peek at the first ten rows of the Law & Order data.

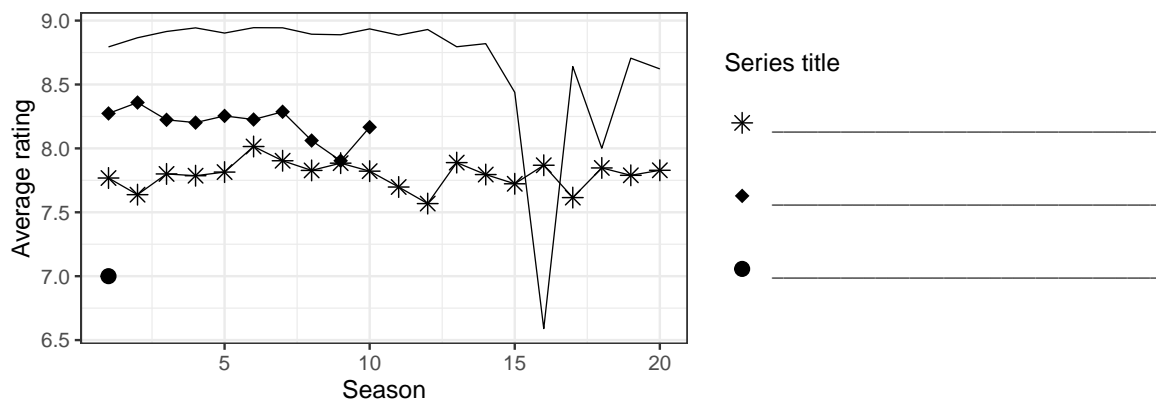
The plot below shows the distributions of average ratings of various Law & Order series across seasons.



Question 8

Based on the information from the side-by-side box plots, fill in the legend of the plot below with Law & Order series titles.

Warning: Removed 21 rows containing missing values or values outside the scale range (`geom_point()`).



Question 9

The following code calculates the standard deviations of average season ratings of the five Law & Order series. Unfortunately, the output is partially erased and replaced with blanks.

```
# A tibble: 5 × 3
  title                                mean_av_rating sd_av_rating
<chr>                                <dbl>         <dbl>
1 Law & Order                        _(1)_         0.106
2 Law & Order: Criminal Intent        8.20         0.129
4 Law & Order: SVU                    8.67         _(2)_
```

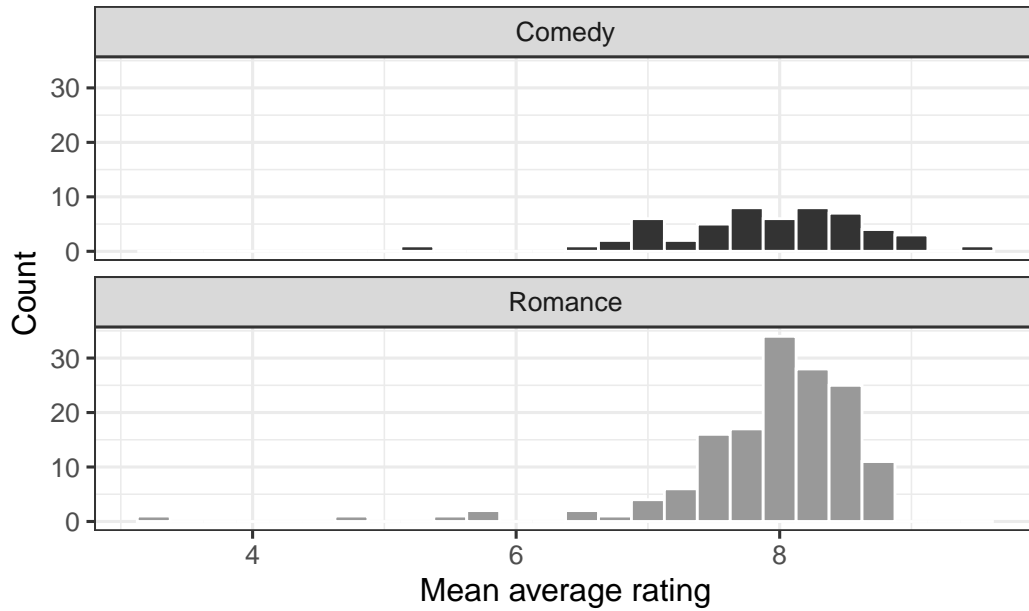
Based on the visualizations you've seen of these data so far, which of the following is true about the blanks in the output? **Select all that are true.**

- a. The **mean** of average ratings (Blank 1) of Law & Order seasons is **lower** than the other two means.
- b. The **mean** of average ratings (Blank 1) of Law & Order seasons is **higher** than the other two means.
- c. The **standard deviation** of average ratings of Law & Order: SVU seasons (Blank 2) is **lower** than the other two standard deviations.
- d. The **standard deviation** of average ratings of Law & Order: SVU seasons (Blank 2) is **higher** than the other two standard deviations.
- e. The **standard deviation** of average ratings of Law & Order: SVU seasons (Blank 2) is **between** the other two standard deviations.

Romance and comedy

Finally, we focus on romance and comedy shows. We first filter the dataset for any shows that have romance or comedy as their genre (**genre_1**, **genre_2**, or **genre_3**) and then remove shows that have both of these genre labels. For the next two questions, we focus on these shows that we identify as either romance or comedy. We then calculate the mean of the average season ratings for each show, to obtain a single “mean average rating” value per show.

The plot below shows the distributions of mean average ratings of seasons of comedy and romance shows.



Question 10

Which of the following statements is **true** about these distributions? **Select all that are true.**

- a. Mean average ratings of romance shows are bimodal.
- b. Mean average ratings of comedy are unimodal.
- c. Mean average ratings of romance shows is left skewed.
- d. Mean average ratings of comedy shows is right skewed.
- e. There are more romance shows than comedy shows.

Data

The data for the next few questions come from the Internet Movie Database (IMDB). Specifically, the data are a random sample of movies released between 1980 and 2020.

The name of the data frame used for this analysis is **movies**, and it contains the variables shown in Table 1.

Table 1: Data dictionary for **movies**

Variable	Description
name	name of the movie

Variable	Description
rating	rating of the movie (R, PG, etc.)
genre	main genre of the movie.
runtime	duration of the movie
year	year of release
release_date	release date (YYYY-MM-DD)
release_country	release country
score	IMDB user rating
votes	number of user votes
director	the director
writer	writer of the movie
star	main actor/actress
country	country of origin
budget	the budget of a movie (some movies don't have this, so it appears as 0)
gross	revenue of the movie
company	the production company

The first thirty rows of the `movies` data frame are shown in Table 2, with variable types suppressed (since we'll ask about them later).

Question 11

The `name` and `runtime` variables are shown below, with the variable types suppressed.

# A tibble: 500 x 2		What is the type of the <code>runtime</code> variable?
name	runtime	
1 Blue City	83 mins	a. Character
2 Winter Sleep	196	b. Double
3 Rang De Basanti	167	c. Factor
4 Pokémon Detective Pikachu	104	d. Integer
5 A Bad Moms Christmas	104	e. Logical
6 Replicas	107	
# i 494 more rows		

Table 2

First 30 rows of movies, with variable types suppressed.

A tibble: 500 x 16

	name	score	runtime	genre	rating	release_country	release_date
1	Blue City	4.4	83 mins	Action	R	United States	1986-05-02
2	Winter Sleep	8.1	196	Drama	Not Rated	Turkey	2014-06-12
3	Rang De Basan~	8.1	167	Comedy	Not Rated	United States	2006-01-26
4	Pokémon Detec~	6.6	104	Action	PG	United States	2019-05-10
5	A Bad Moms Ch~	5.6	104	Comedy	R	United States	2017-11-01
6	Replicas	5.5	107	Drama	PG-13	United States	2019-01-11
7	Windy City	5.8	103	Drama	R	Uruguay	1986-01-01
8	War for the P~	7.4	140	Action	PG-13	United States	2017-07-14
9	Tales from th~	6.4	98	Crime	R	United States	1995-05-24
10	Fire with Fire	6.5	103	Drama	PG-13	United States	1986-05-09
11	Raising Helen	6	119	Comedy	PG-13	United States	2004-05-28
12	Feeling Minne~	5.4	99	Comedy	R	United States	1996-09-13
13	The Babe	5.9	115	Biography	PG	United States	1992-04-17
14	The Real Blon~	6	105	Comedy	R	United States	1998-02-27
15	To vlemma tou~	7.6	176	Drama	Not Rated	United States	1997-11-01
16	Going the Dis~	6.3	102	Comedy	R	United States	2010-09-03
17	Jung on zo	6.8	103	Action	R	Hong Kong	1993-06-24
18	Rita, Sue and~	6.5	93	Comedy	R	United Kingdom	1987-05-29
19	Phone Booth	7	81	Crime	R	United States	2003-04-04
20	Happy Death D~	6.6	96	Comedy	PG-13	United States	2017-10-13
21	Barely Legal	4.7	90	Comedy	R	Thailand	2006-05-25
22	Three Kings	7.1	114	Action	R	United States	1999-10-01
23	Menace II Soc~	7.5	97	Crime	R	United States	1993-05-26
24	Four Rooms	6.8	98	Comedy	R	United States	1995-12-25
25	Quartet	6.8	98	Comedy	PG-13	United States	2013-03-01
26	Tape	7.2	86	Drama	R	Denmark	2002-07-12
27	Marked for De~	6	93	Action	R	United States	1990-10-05
28	Congo	5.2	109	Action	PG-13	United States	1995-06-09
29	Stop-Loss	6.4	112	Drama	R	United States	2008-03-28
30	Con Air	6.9	115	Action	R	United States	1997-06-06
	budget	gross	votes	year	director	writer	star
1	10000000	6947787	1100	1986	Michelle Manning	Ross Macdona~	Judd Nelson
2	NA	4018705	48000	2014	Nuri Bilge Ceyl~	Ebru Ceylan	Haluk Bilgin~
3	NA	10800778	115000	2006	Rakeysh Ompraka~	Renzil D'Sil~	Aamir Khan
4	150000000	433921300	146000	2019	Rob Letterman	Dan Hernandez	Ryan Reynolds
5	28000000	130560428	46000	2017	Jon Lucas	Jon Lucas	Mila Kunis
6	30000000	9330075	34000	2018	Jeffrey Nachman~	Chad St. John	Keanu Reeves
7	NA	343890	262	1984	Armyan Bernstein	Armyan Berns~	John Shea
8	150000000	490719763	235000	2017	Matt Reeves	Mark Bombback	Andy Serkis
9	6000000	11837928	7400	1995	Rusty Cundieff	Rusty Cundie~	Clarence Wil~
10	NA	4636169	1500	1986	Duncan Gibbins	Bill Phillips	Craig Sheffer
11	50000000	49718611	36000	2004	Garry Marshall	Patrick J. C~	Kate Hudson
12	NA	3124440	11000	1996	Steven Baigelman	Steven Baige~	Keanu Reeves
13	NA	19930973	9300	1992	Arthur Hiller	John Fusco	John Goodman
14	NA	83488	3900	1997	Tom DiCillo	Tom DiCillo	Matthew Modi~
15	NA	NA	6400	1995	Theodoros Angel~	Theodoros An~	Harvey Keitel
16	32000000	42059111	57000	2010	Nanette Burstein	Geoff LaTuli~	Drew Barrymo~

Question 12

The code below summarizes the data in a certain way.

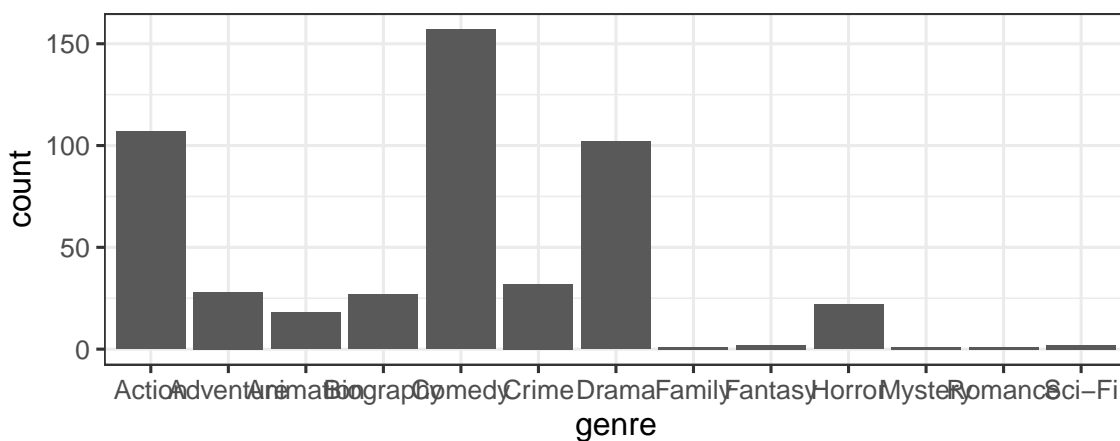
```
# A tibble: 1 x 1
  `sum(release_country == "United States")`
    <int>
1                435
```

Which of the following is **TRUE** about the code and its result? **Select all that are true.**

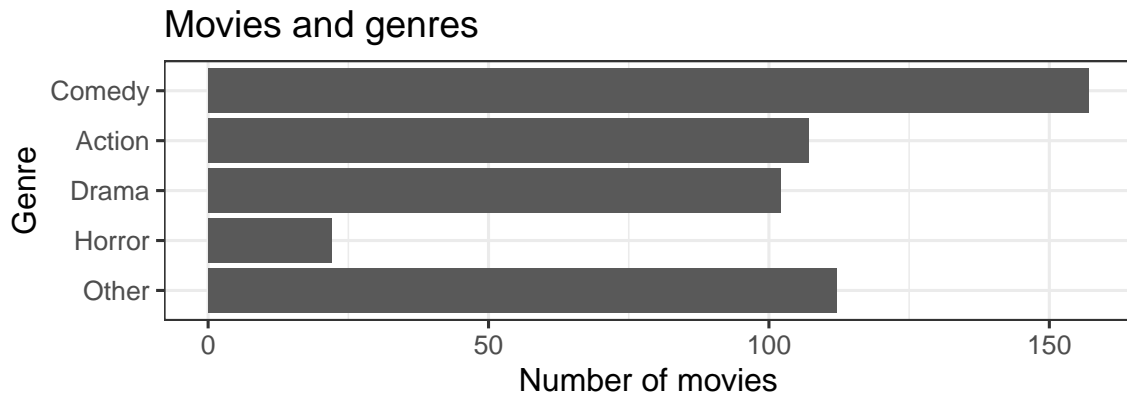
- a. Evaluates whether each `release_country` is equal to "United States" or not, which results in a logical variable.
- b. Filters out rows where `release_country` is not equal to "United States" and counts the remaining rows.
- c. Sums the logical values, where each TRUE is considered a 1 and each FALSE is considered a 0.
- d. Results in a character vector.
- e. The result shows there are 435 movies released in the United States.

Question 13

Suppose you want a visualization that shows the number of movies in the sample in each `genre`. Your first attempt is as follows.



A friend of yours says that the visualization is difficult to read and they suggest using the following visualization instead.



Source: IMDB.

Which of the following modifications would your friend have made to your code to create their version? **Select all that apply.**

- a. Combine movies in genres other than Comedy, Drama, Action, and Horror into a new level called "Other".
- b. Reorder the levels in descending order of numbers of observations, except for the "Other" level.
- c. Map **genre** to the y aesthetic.
- d. Add a title, x and y-axis labels, and a caption.
- e. Filter out all moves in genres other than Comedy, Drama, Action, and Horror before plotting.

Question 14

Which of the following is **TRUE** about the code and its result? **Select all that are true.**

```
# A tibble: 6 x 6
  rating    Other Drama Action Comedy Horror
  <fct>    <int> <int> <int> <int> <int>
1 G         5     1     1     1     0
2 PG        38    13    10    18     0
3 PG-13     19    25    35    35     0
4 R         45    50    57    96    21
5 NC-17      1     2     0     1     0
6 Not Rated  4    11     4     6     1
```

- a. The code counts how many movies are in each rating and genre combination.

- b. The code sorts the results in descending order.
- c. Each row of the output is a movie.
- d. The output shows that there are six distinct ratings in the dataset.
- e. The code reduces the number of variables and observations in the `movies` data frame to six.