ACTIVITY 4

IMDb Movie Reviews

4.1 Learning objectives

- Identify and create appropriate summary statistics and plots given a data set or research question for quantitative data
- Interpret the following summary statistics in context: median, lower quartile, upper quartile, standard deviation, inter-quartile range
- Given a plot or set of plots, describe and compare the distribution(s) of a single quantitative variable (center, variability, shape, outliers)

4.2 Terminology review

In today's activity, we will review summary measures and plots for quantitative variables. Some terms covered in this activity are:

- Two measures of center: mean, median
- Two measures of spread (variability): standard deviation, inter-quartile range (IQR)
- Types of graphs: box plots, dot plots, histograms

To review these concepts, see Section 2.3 in the textbook.

4.3 Movies released in 2016

A data set was collected on movies released in 2016. Here is a list of some of the variables collected on these movies.

Variable	Description
`budget_mil`	Amount of money (in US $ millions) budgeted for the production of the movie
`revenue_mil`	Amount of money (in US $ millions) the movie made after release
`duration`	Length of the movie (in minutes)
`content_rating`	Rating of the movie (`G`, `PG`, `PG-13`, `R`, `Not Rated`)
`imdb_score`	IMDb user rating score from 1 to 10
`genres`	Categories the movie falls into (e.g., Action, Drama, etc.)
`movie_facebook_likes`	Number of likes a movie receives on Facebook

Vocabulary review

1. What are the observational units in this data set?

2. Which of the above listed variables are categorical?

3. Which of the above listed variables are quantitative?

Summarizing a single quantitative variable

The `favstats` function gives the summary statistics for a quantitative variable. Here we have the summary statistics for the variable `imdb_score`.

```
movies <- read.csv("data/Movies2016.csv") # Read in data set
movies %>% #Data set piped into...
  summarise(favstats(imdb_score)) #Apply favstats function to imdb_score
```

```
#>   min   Q1 median  Q3 max     mean       sd  n missing
#> 1 3.4 5.65    6.4 7.1 8.2 6.309783 1.086689 92       0
```

4. Give the values for the two measures of center.

5. Calculate the IQR.

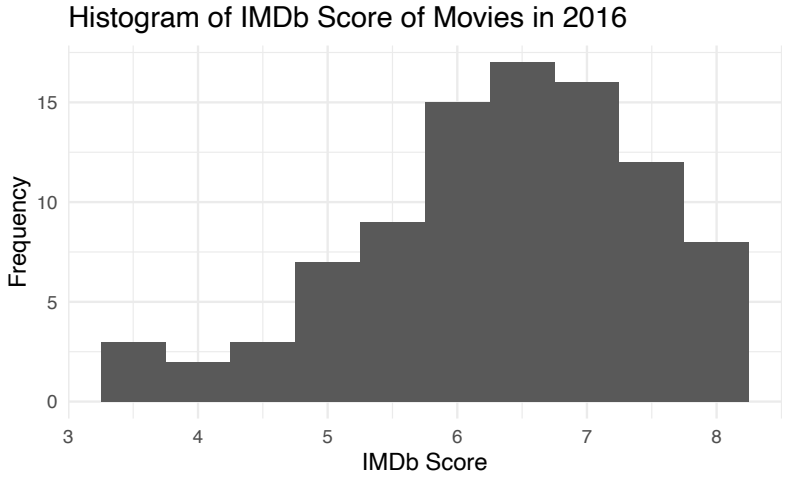
6. Report the value of the standard deviation and interpret this value in context of the problem.

Displaying a single quantitative variable

7. What are the three types of plots used to plot a single quantitative variable?

A histogram of the IMDb scores is shown below. Visually, this shows us the range of IMDb scores for Movies released in 2016. Notice that the **bin width** is 0.5. For example the first bin consists of the number of movies in the data set with an IMDb score of 3.25 to 3.75. It is important to note that a movie with a IMDb score on the boundary of a bin will fall into the bin above it; for example, 4.76 would be counted in the bin 4.75–5.25.

```
movies %>% #Data set piped into...
ggplot(aes(x = imdb_score)) +   #Name variable to plot
  geom_histogram(binwidth = 0.5) +  #Create histogram with specified binwidth
  labs(title = "Histogram of IMDb Score of Movies in 2016", #title for plot
       x = "IMDb Score", #Label for x axis
       y = "Frequency") #Label for y axis
```



8. Which range of IMDb scores have the highest frequency?

9. What is the shape of the distribution of IMDb scores?

10. Which five summary statistics are used in creating a box plot? *Hint*: Together they are called the **five-number summary** of the variable.

11. The three smallest IMDb scores in the data set are 3.4, 3.5, and 3.7 and the three largest IMDb scores are 8.5, 8.7, and 9.1:

```
movies %>% # Data set pipes into...
  select(imdb_score) %>% # Select imdb_score variable
  slice_min(imdb_score, n = 3)  # Show 3 smallest values
```

```
#>   imdb_score
#> 1        3.4
#> 2        3.5
#> 3        3.7
```

```r
movies %>% # Data set pipes into...
  select(imdb_score) %>% # Select imdb_score variable
  slice_max(imdb_score, n = 3)  # Show 3 largest values
```

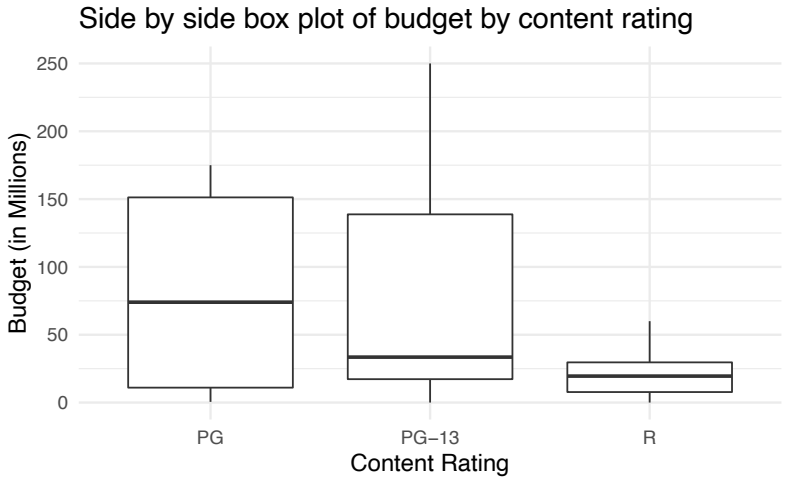
```
#>   imdb_score
#> 1        8.2
#> 2        8.1
#> 3        8.0
```

Using the summary statistics above, and the smallest and largest values of variable to check for outliers, sketch a box plot of IMDb Score. Be sure to label the axes.

Displaying a single categorical and single quantitative variable

The box plot of movie budgets (in millions) by content rating is plotted using the code below. This plot helps to compare the budget for different levels of content rating.

```r
movies %>%  #Data set piped into...
  filter(content_rating != "Not Rated") %>% # Remove Not Rated movies
  ggplot(aes(y = budget_mil, x = content_rating))+  #Identify variables
  geom_boxplot()+  #Tell it to make a box plot
  labs(title = "Side by side box plot of budget by content rating",  #Title
       x = "Content Rating",    #x-axis label
       y = "Budget (in Millions)")  #y-axis label
```



12. Answer the following questions about the box plots above.

a. Which content rating has the highest center?

b. Which content rating has the largest spread?

c. Which content rating is the most symmetric distribution?

d. Fifty percent of movies in 2016 with a PG-13 content rating fall below what value?

e. What is the value for the third quartile (Q3) for the PG-13 rating? Interpret this value in context.

4.4 Additional notes

Use this space to summarize your thoughts and take additional notes on today's activity.