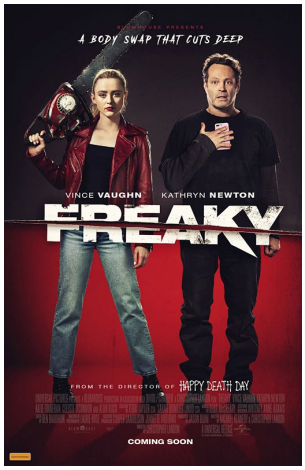
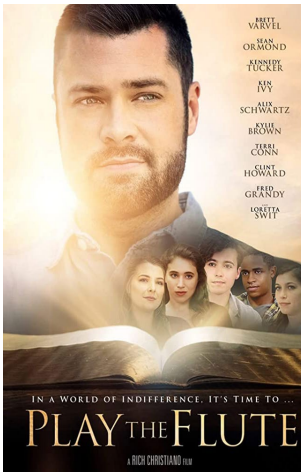
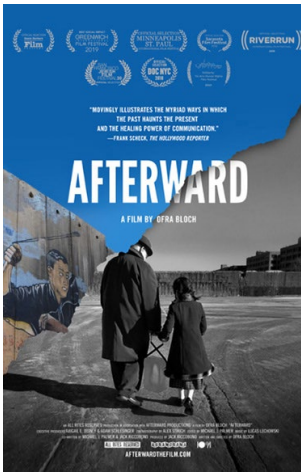
Exploring Quantitative Variables: IMDb Movie Reviews

Your Name: Dr. Theobold's Key

September 27, 2022







Learning outcomes

- 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, first quartile, third quartile, standard deviation, interquartile range.
- Identify and create appropriate summary statistics and plots given a data set or research question for a single quantitative variable.
- Given a plot or set of plots, describe and compare the distribution(s) of a single quantitative variable (center, variability, shape, outliers).

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, interquartile range (IQR)
- Types of graphs: box plots, dot plots, histograms

Movies released in 2016

A data set was collected on movies released in 2020. Here is a list of some of the variables collected on the observational units (each movie):

Variable	Description
`Movie`	Title of the movie
`averageRating`	Average IMDb user rating score from 1 to 10
`numVotes`	Number of votes from IMDb users
`Genre`	Categories the movie falls into (e.g., Action, Drama, etc.)
`2020 Gross`	Gross profit from movie viewing
`runtimeMinutes`	Length of movie (in minutes)

Visualizing a Single Quantitative Variable

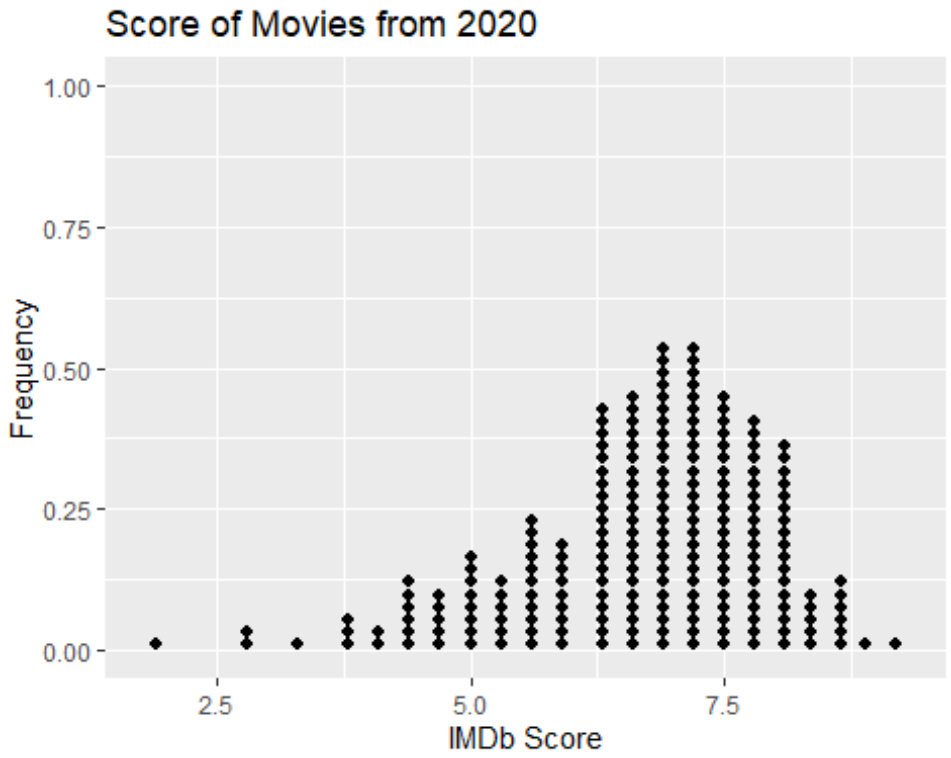
1. What are the three types of plots used to plot a single quantitative variable?

- Boxplot
- Histogram
- Dotplot

Dotplot

A dotplot will plot a dot for each value in the data set. The code below was used to create a dotplot of the `averageRatings` variable from the `movies` dataset. In a dotplot, the quantitative variable goes on the x-axis, which is why the code says `x = averageRating` inside of the `aes()` function.

```
ggplot(data = movie_ratings,
       mapping = aes(x = averageRating)) +
  geom_dotplot(dotsize = 0.5) +
  labs(title = "Score of Movies from 2020", # Title for plot
       x = "IMDb Score", # Label for x axis
       y = "Frequency" # Label for y axis
       )
```



2. What does each dot on the dotplot represent?

   Each dot is one movie, or the IMDb score for one movie.

3. How would you describe the shape of the distribution of IMDb scores?

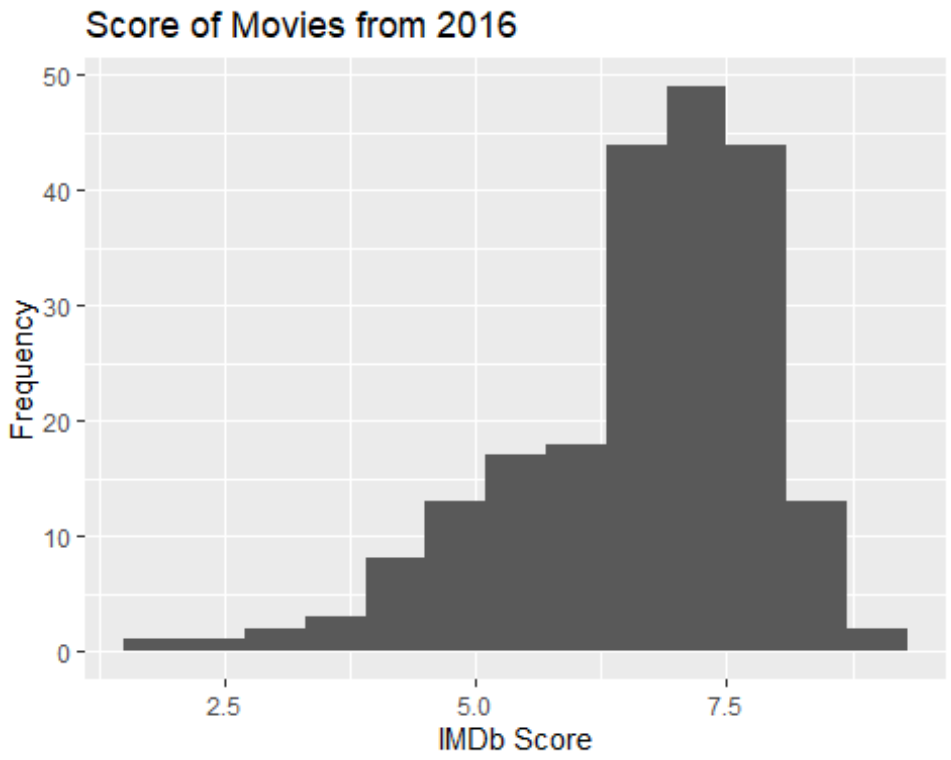
   The shape is unimodal (one peak) and left skewed (long left tail).

Histogram

To create a histogram of the IMDb scores, all we need to do is change the geometric object we are displaying on our plot! In a dotplot we use dots, but in a histogram we use bars. Notice, in the code below there are two changes:

- I am using `geom_histogram()` instead of `geom_dotplot()`
- I am specifying how wide the bins of the histogram should be using `binwidth = 0.6`

```
ggplot(data = movie_ratings,
       mapping = aes(x = averageRating)) +
  geom_histogram(binwidth = 0.6) +
  labs(title = "Score of Movies from 2016", # Title for plot
       x = "IMDb Score", # Label for x axis
       y = "Frequency" # Label for y axis
       )
```



4. Why did I **not** need to specify a `binwidth` in the dot plot I made?

   Because each dot was plotted on its own!

5. Which range of IMDb scores have the *highest* frequency?

   Movie scores between 6 and 8.

5. What IMDB scores are movies *rarely* rated?

   Scores less than 4 seem rare.

6. Are there IMDB scores that were possible but *no* movies in this sample were given those ratings?

   Yes! Movies can score up to 10 and as low as 0. This sample didn't have ratings above 9 or below 2.

Boxplot

7. Which five summary statistics are used to create a box plot?

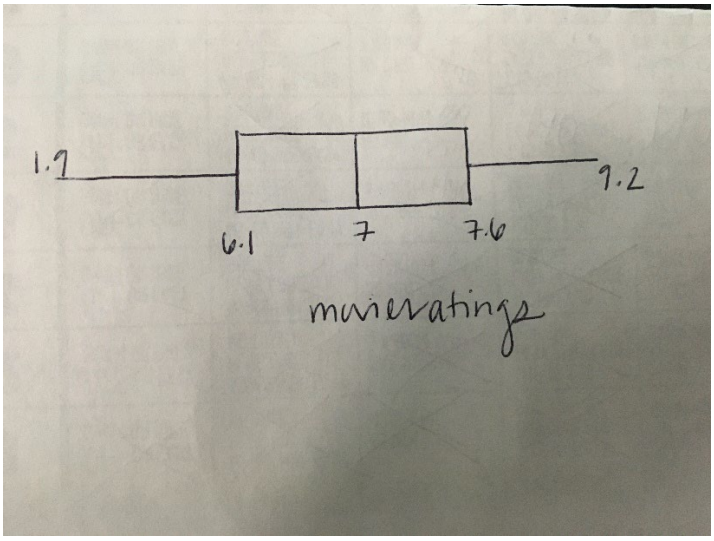
   - Minimum
   - Quantile 1 (the 25th percentile)
   - Median
   - Quantile 3 (the 75th percentile)
   - Maximum

In the code below I've provided you with all of the statistics you listed in #7.

```
summarize(movie_ratings,
          min_score = min(averageRating),
          Q1_score = quantile(averageRating, 0.25),
          median_score = median(averageRating),
          Q3_score = quantile(averageRating, 0.75),
          max_score = max(averageRating)
          )

# A tibble: 1 × 5
  min_score Q1_score median_score Q3_score max_score
      <dbl>    <dbl>        <dbl>    <dbl>     <dbl>
1       1.9      6.1            7      7.6       9.2
```

8. Using the summary statistics provided, sketch a box plot of IMDb scores. Be sure to label the axes!



9. How do you decide if a value is an "outlier" when creating a boxplot?

   You check if the value is outside the "fences," which are calculated as Q1 - 1.5xIQR and Q3 + 1.5xIQR. Any value lower than the lower fence or above the upper fence is considered an outlier.

In the code below, I'm providing you with the top 3 and the bottom 3 IMDb scores.

Bottom 3:

```
movie_ratings %>%
  select(averageRating) %>%
  slice_min(order_by = averageRating, n = 3)

# A tibble: 3 × 1
  averageRating
          <dbl>
1           1.9
2           2.7
3           2.9
```

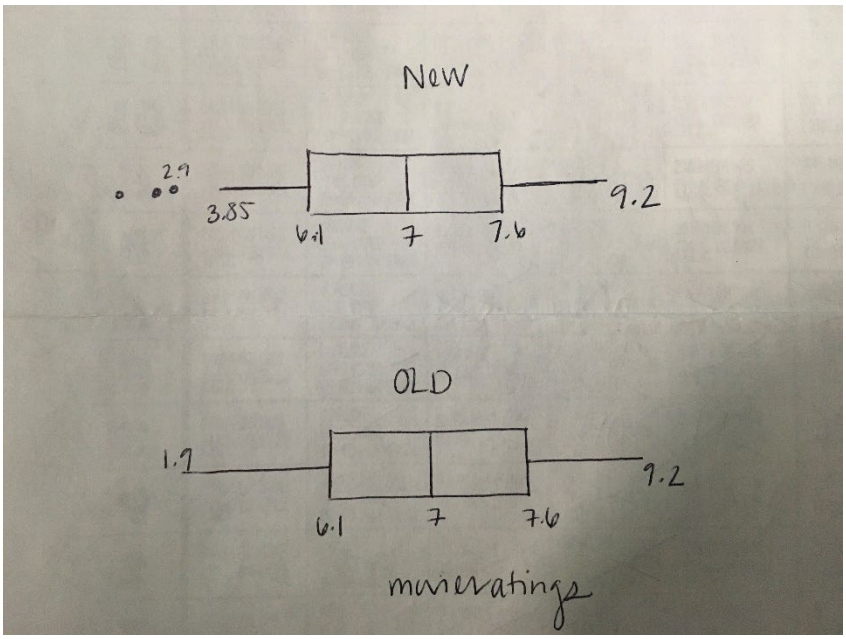
Top 3:

```
movie_ratings %>%
  select(averageRating) %>%
  slice_max(order_by = averageRating, n = 3)

# A tibble: 6 × 1
  averageRating
          <dbl>
1           9.2
2           8.9
3           8.7
4           8.7
5           8.7
6           8.7
```

10. Revisit your previous boxplot to decide if any observations should be plotted as outliers. (Modify your previous plot)

The upper end of the plot didn't change, since the upper fence is 9.85. The lower end did change, since the three lowest values are below the lower fence (3,85).



Plot Comparison

11. Compare the three graphs of IMDb scores created above.

- Which graph(s) show the shape of the distribution?

  Both the dotplot and the histogram show the shape.

- Which graph(s) show the outliers of the distribution?

  The boxplot is the easiest for seeing outliers, but they are still noticeable in the dotplot and histogram.

- Which graph plots the *raw* data (individual observations)?

  Only the dotplot shows the individual observations.

Summarizing a single quantitative variable

12. Based on the distributions provided, do you believe the *mean* IMDb score will be greater or less than the median? Explain why!

In the code below I've calculated the standard deviation of the IMDb scores.

```
summarize(movie_ratings,
          sd_score = sd(averageRating)
          )

# A tibble: 1 × 1
  sd_score
```

```
     <dbl>
1     1.25
```

13. Interpret the value of the standard deviation in the context of these data.

    We expect that ***most*** of the movie ratings to fall about 1.25 away from the mean movie rating.