Quantitative Exploratory Data Analysis

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)

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
  - IQR
- Boxplots, dotplots, histograms

To review these concepts see Section 2.3 in the textbook.

Movies Released in 2016

A data set was collected on Movies released since 1916 to 2016. Here is a list of some of the variables collected on these movies.

- Year: Year the movie was released
- Budget: The amount of money (in US $ millions) budgeted for the production of the movie
- Revenue: The amount of money (in US $ millions) the movie made after release
- Duration: The length of the movie (in minutes)
- Content Rating: Rating of the movie (G, PG, PG-13, R, Not Rated)
- IMDb Score: User rating score from 1 to 10
- Genre: Category the movie falls into
- 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'.

```
favstats(moviesa$imdb_score)
```

```
##  min  Q1 median  Q3 max     mean        sd  n missing
##  3.4 5.9    6.6 7.1 8.2 6.459016 0.9218418 61       0
```

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

5. Calculate the IQR.

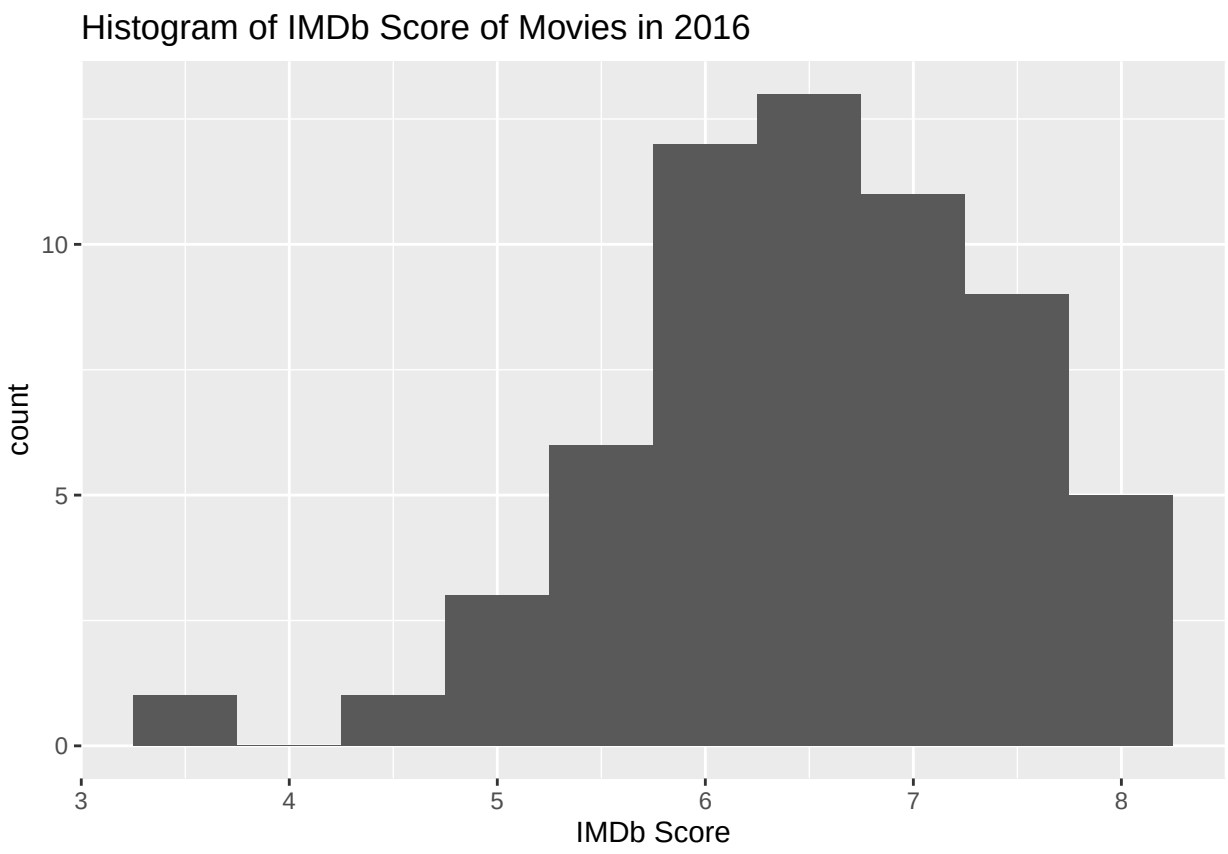
6. Report the value of the standard deviation and interpret this value in context of the problem.

Displaying a single quantitative variable

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

A histogram of the variable 'IMBd Score' is shown below. 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 IMBd score of 3.25 to 3.75. It is important to note that a movie with a IMBd score of 4.75 will fall into the bin for 4.75 - 5.25. Visually this shows us the range of IMBd scores for Movies released in 2016.

```
ggplot(data = moviesa,    #Name data set
       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
```



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

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

The boxplot is created using the five number summary:

- Minimum value
- Quartile 1 (Q1) - the value at the 25th percentile
- Median - the value at the 50th percentile
- Quartile 3 (Q3) - the value at the 75th percentile

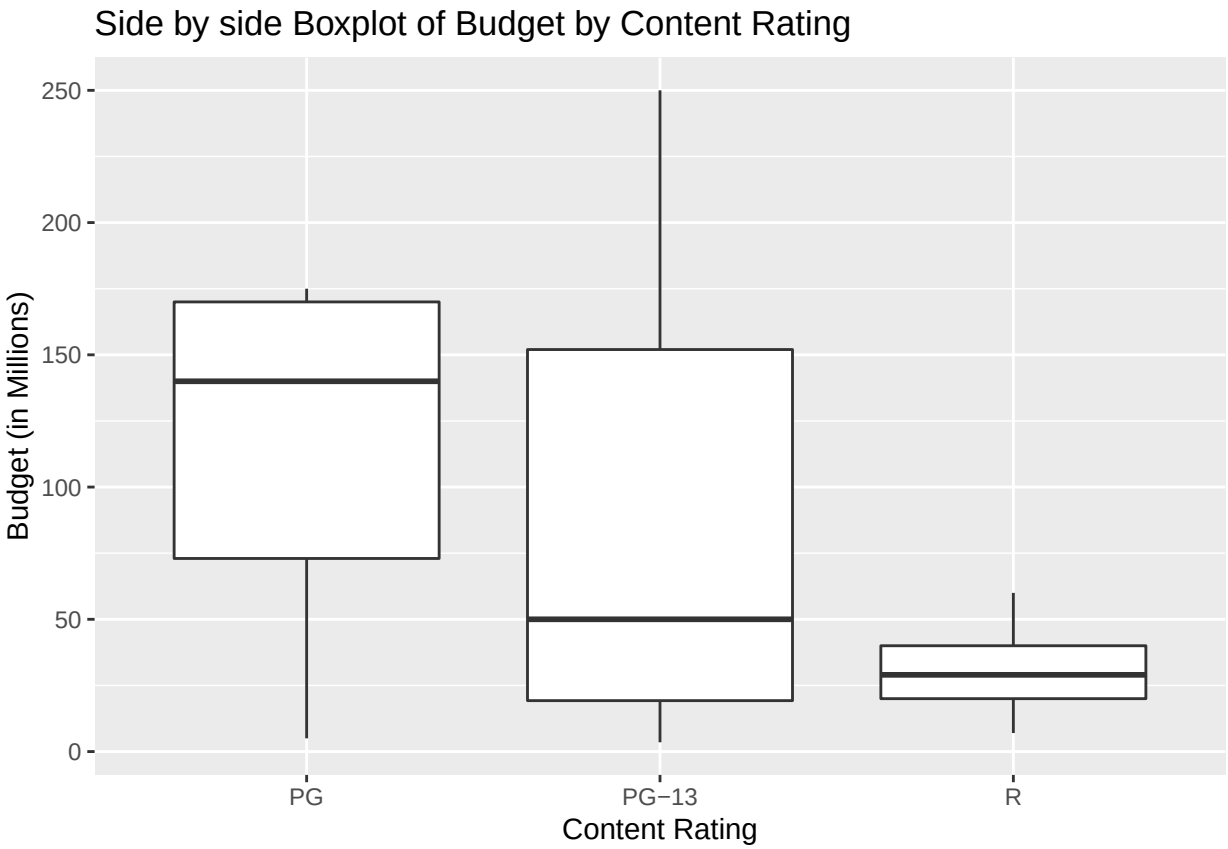
- Maximum value

10. 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. Using the summary statistics above, sketch a boxplot of IMDb Score. Be sure to label the axes.

Displaying a Single Categorical and Single Quantitative Variable

The boxplot of 'Budget' in millions by 'Content rating' is plotted using the code below. This plot helps to compare the budget for different levels of content rating.

```
ggplot(data = moviesa,  #Data set
       aes(y = budget_mil, x = content_rating))+  #Identify variables
  geom_boxplot()+  #Tell it to make a boxplot
  labs(title = "Side by side Boxplot of Budget by Content Rating",  #Title
        x = "Content Rating",    #x-axis label
       y = "Budget (in Millions)")  #y-axis label
```



11. Answer the following questions about the boxplots 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?

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

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