Project 1

Due: Wednesday April 20th, 2016 at 1pm

Project should be turned into the stat 67 box (DBH 2nd floor) or in class.

The point of the project is to produce graphics and summary statistics you would be proud to show in a presentation.

Download the ANES data set and ANES codebook from the class website. You will need to choose 1-2 categorical variables and 2-3 numeric variables you find interesting. For this data set, almost all the variables are categorical even thought they are presented as numeric. You will need to relabel the levels of the variables with their proper names when you present statistics or graphics. More details can be found in the codebook.

The project must be printed and stapled in the upper left hand corner with name in the upper right hand corner. Each plot/table must be printed at a readable size and text must be next to, above, or below plot. Do not place all the plots/tables or all text at the end.

Each question will be graded in the following way (see the end of this document for some examples):

- Graphic is appropriate to the data type / Statistics are meaningful and appropriate to data type (3 points)
- Title or Label (1 point)
- Clearly labeled axes/key (2 point)
  - labels on axes, keys and tables are words that are interpretable by someone not looking at the codebook
- Appropriate scale/cleaned data (2 point)
  - Scale is meaningful (points are not all scrunched up, enough tick marks to read the plot, etc.)
  - Missing values removed or placed in their own category
  - Mislabeled levels are combined
  - levels are properly ordered (if ordinal or numeric)
- A one sentence (complete sentence) description of what your graphic/statistic shows (2 point)

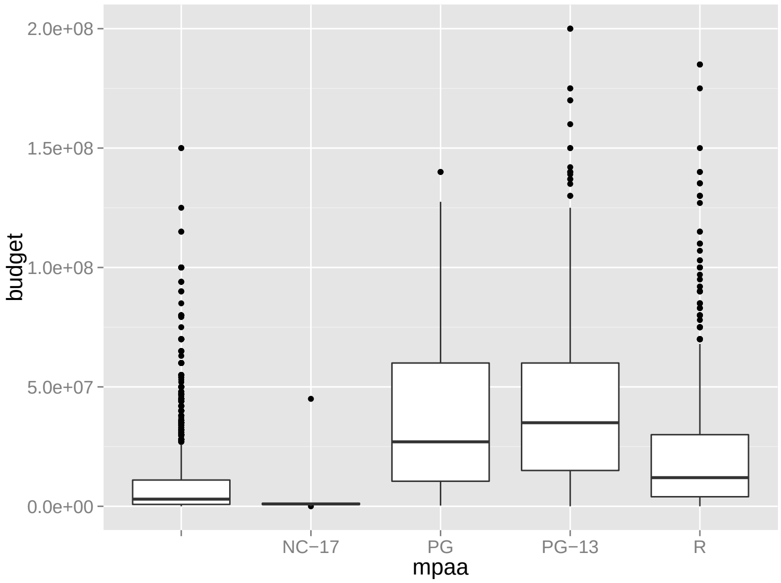
1. Calculate summary statistics for any one numeric variable. Write a one to two sentence explanation of your statistics.
2. Create a statistical graphic appropriate for the numeric variable. Write a one or two sentence explanation of of your graphic.
3. Turn your variable into a categorical variable using the cut function (you may also want the quantile function). You may name your categories with words (eg. high, medium, low) or as ranges. Choose what you feel is an appropriate number of categories. Calculate summary statistics appropriate for the categorical variable. Write a one or two sentence explanation of your statistics.
4. Create a statistical graphic appropriate for the categorical variable. Write a one or two sentence explanation of of your graphic.
5. Choose any two categorical variables and create a statistical graphic showing their relationship. Write a one or two sentence explanation of the graphic in the context of the variables.

6. Choose any two numeric variables and create a statistical graphic showing their relationship. Write a one or two sentence explanation of the graphic in the context of the variables.

7. Choose any categorical variable and any numeric variable. Create a statistical graphic showing their relationship. Write a one or two sentence explanation of the graphic in the context of the variables.

8. Create a graphic of any 3 variables by adding size, color, or groups to your plot. Write a one or two sentence explanation of the graphic in the context of the variables.

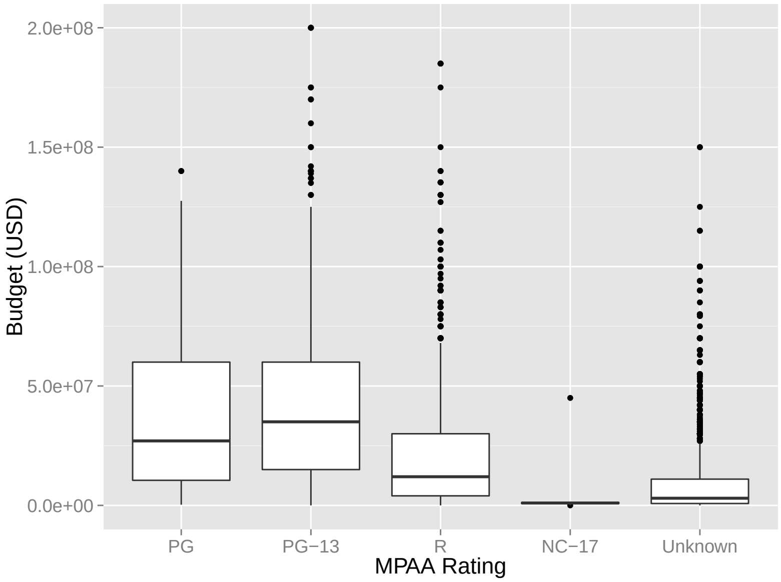
1 Some Examples of Bad and Improved Plots



Budget differs by MPAA rating; NC-17 and R rated films are more likely to have lower budgets.

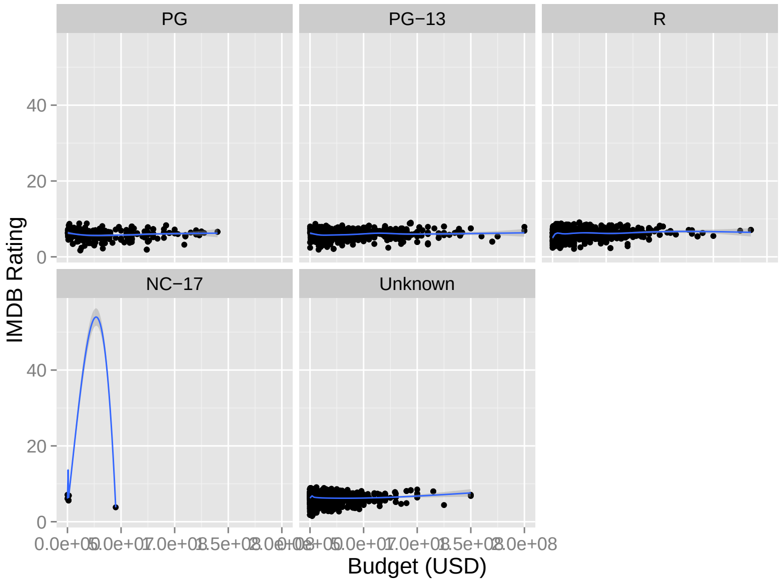
This plot has no title or label. The MPAA rating has levels which are not ordered properly. The units on budget are unknown. "mpaa" is not a very meaningful axis label. There is a level with no label.

Figure 1: MPAA Rating and Budget

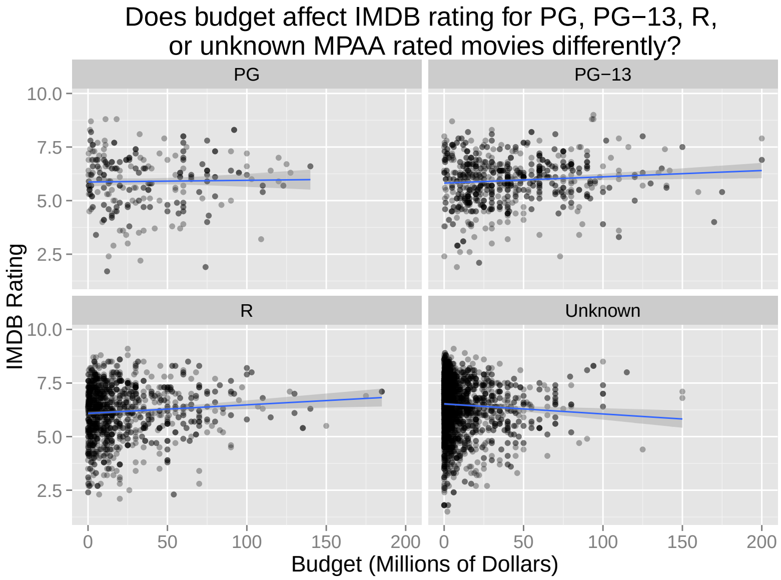


In Figure 1, we can see that budget differs by MPAA rating. NC-17 and R rated films are more likely to have lower budgets.

This plot is improved. It now has a label, the label is referenced in the text. The levels are properly ordered, and labeled. The axes are more meaningful.



This plot has no title or label. The points are all scrunched up because the scale is bad. In addition, the NC-17 rated films do not look meaningful. The x-axis has unreadable scales.



This plot is improved. It now has a title. The points are not all scrunched up. The NC-17 box has been removed for clarity. The x-axis has a readable, meaningful scale. The points are also semi-transparent so that regions of higher density are clearly visible.

In PG, PG-13 and R rated movies, IMDB rating is likely to increase slightly as budget increases. When the rating is unknown, there appears to be a decreasing trend, but it could be due to outliers.