MTH208a: Worksheet 11

Visualizing Data

So far, we have learned how to collect data, clean and process it, and save it. Through courses in the next few years, you will learn how to analyze the data through statistical models.

However, a crucial component of data analysis is data visualization. This can often help ask interesting questions about the data.

We can think of visualizations as:

- single variable visualization - histogram, boxplots
- Multivariable visualization - scatterplot, side-by-side boxplot.

Using the movies data-set scraped in Worksheet 7 we will try to visualize the data.

Note: In any visualization one must be very clear about what we are trying to visualize. Further axes should be clearly described and legends should be

1. You can find the data in `IMDB_movies.Rdata` in your GitHub repository. Download the `.Rdata` in your working directory and load the file using the `load()` function.

2. **Histogram**: read the documentation for the `hist()` function that makes a histogram.
   a. Make a histogram of the ratings for the top 250 movies. Using the argument `main`, set the title of the histogram to be "Histogram of Ratings" and the label on the x-axis as "Ratings".
   b. Make the histogram again so that the bars are white in color.

3. **Boxplot**: read the documentation for the `boxplot()` function that makes a boxplot. The line in the middle is the median of the observations, the bottom box the 25\% percentile, the top box the 75\% percentile.
   a. Make a boxplot of the ratings of the top 250 movies. Make sure to assign an appropriate title.
   b. Make the boxplot again so that the bars are pink in color.

4. **Side-by-Side Boxplots:** Reading the help page for `boxplot()`, make a side-by-side boxplot of men's ratings and women's ratings. Make sure the axis labels are appropriate.

5. **Overlapped histograms:** Make a plot of histograms of men's ratings and women's ratings, overlayed on top of each other. You may use the `col = adjustcolor("red", alpha.f = .5)` option to make colors transparent.

Use `legend()` command to add a legend to the plot.

6. **Scatterplot:** Scatterplots can help explain the relationship between two quantitative variables. Using the `plot()` function, make a scatterplot of number of votes on the x-axis and the ratings on the y-axis.

7. **Text:** Using the `text()` function, write down the names of the movies in the above plot whose ratings are more than 8.9.

8. **3 variable plots in two dimensions:** Visualizing a third variable in the above plot is possible when the third variable is categorical.
   IMDB.com was established in the year 1996. In the plot in Part (6), color the movies released before 1996 in a different color from the movies released after 1996. Make sure to add a legend.

9. Using such visualizations can you identify some potential biases or unfairness in the movie rankings?

10. Create an animation using `Sys.sleep()` and `points()` function, of presenting one data-point at a time in Park (8).