DLI Accelerated Data Science Teaching Kit

Module 9 Lab:
Web-based Data Visualization (D3)

Q1 [15 points] Scatter plots

Use the dataset[1] provided in the file *movies.csv* to create a scatter plot.

Refer to the tutorial for scatter plot here.

Attributes in the dataset:
- Feature 1: Id of the movie
- Feature 2: Title
- Feature 3: Year
- Feature 4: Runtime (minutes)
- Feature 5: Country
- Feature 6: IMDb Rating
- Feature 7: IMDb Votes
- Feature 8: Budget (in USD)
- Feature 9: Gross (in USD)
- Feature 10: Wins and nominations
- Feature 11: Is good rating? (value 1 means "good", value 0 - "bad")

Optional: to learn more about IMDb, visit https://en.wikipedia.org/wiki/IMDb

a. [8 points] Creating scatter plots:

1. **[6 points] Create two scatter plots**, one for each feature combination specified below. In the scatter plots, visualize "good rating" class instances as blue crosses, and "bad rating" instances as red circles. Add a legend to the top right corner showing the symbols' mapping to the classes.
   - Feature 10 (Wins and nominations) vs. Feature 6 (IMDb Rating)
     - Figure title: Wins+Nominations vs. IMDb Rating
     - X axis (horizontal) label: IMDb Rating
     - Y axis (vertical) label: Wins+Noms
   - Feature 8 (Budget) vs. Features 6 (IMDb Rating)

[1] Source: derived from a "movies" dataset prepared by Dr. Guy Lebanon, for an earlier version of OMSCS CSE 6242 (the source raw data is available at the following URL; you do not need to download it when working on this question https://s3.amazonaws.com/content.udacity-data.com/courses/gt-cs6242/project/movies_merged)

- Figure title: Budget vs. IMDb Rating
- X axis (horizontal) label: IMDb Rating
- Y axis (vertical) label: Budget

2. **[2 points]** In **explanation.txt**, use no more than 50 words to discuss which feature combination is better at separating the classes and why.

Note: Your scatter plots should be placed one after the other **on a single HTML page**, similar to the example image below. Note that your design need NOT be identical to the example.

b. [3 points] Scaling symbol sizes. Create a scatter plot (append to the HTML page) using the feature combination specified below. Set the size of each symbol to be proportional to the value of Feature 10 (Wins and nominations); use a good scaling coefficient to make the scatter plot legible, visually attractive and meaningful. Visualize "good rating" class instances as blue crosses, and "bad rating" instances as red circles.

- Feature 7 (IMDb Votes) vs. Feature 6 (IMDb Rating) sized by Feature 10 (Wins+Nominations)
  - Figure title: Votes vs. IMDb Rating sized by Wins+Nominations
  - X axis (horizontal) label: IMDb Rating
  - Y axis (vertical) label: IMDb Votes

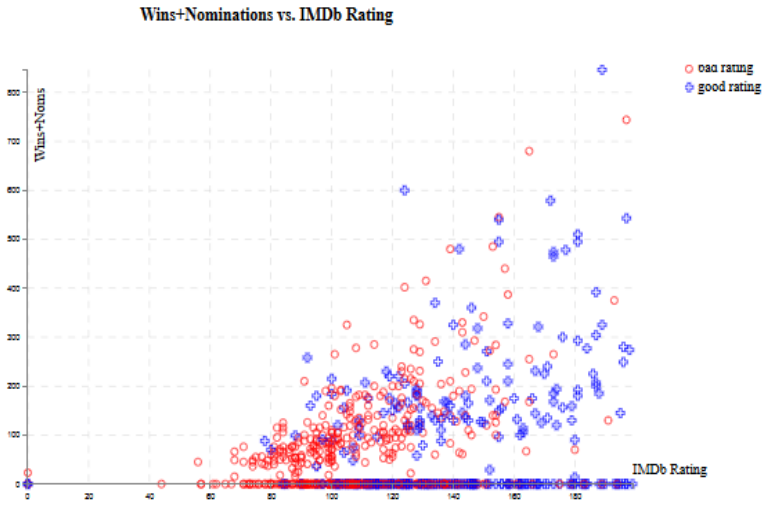
c. [4 points] Axis scales in D3. Create two plots for this part (append to the HTML page) to try out two axis scales in D3: the first plot uses the square root scale for its y-axis (only), and the second plot uses the log scale for its y-axis (only). In **explanation.txt,** explain when we may want to use square root scale and log scale in charts, in no more than 50 words.

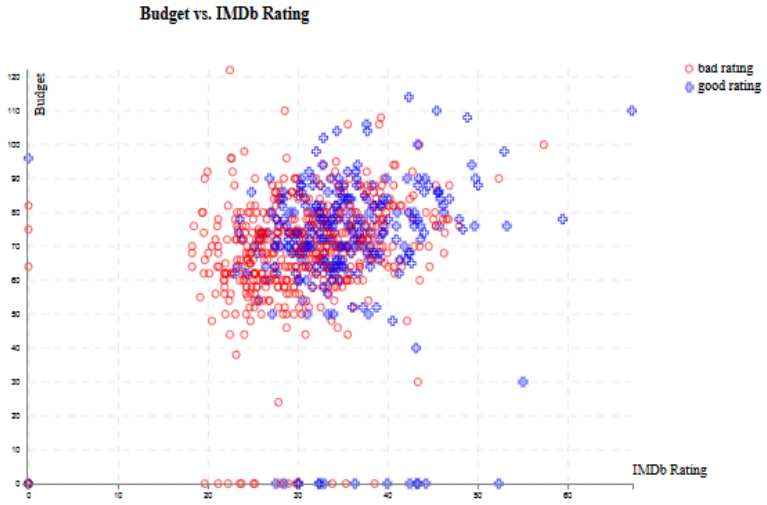
Note: the x-axes should be kept in linear scale, and only the y-axes are affected.

Hint: You may need to carefully set the scale domain to handle the 0s in data.

- First Figure: uses the square root scale for its y-axis (only)
  - Figure title: Wins+Nominations (square-root-scaled) vs. IMDb Rating
  - X axis (horizontal) label: IMDb Rating
  - Y axis (vertical) label: Wins+Noms

- Second Figure: uses the log scale for its y-axis (only)
  - Figure title: Wins+Nominations (log-scaled) vs. IMDb Rating
  - X axis (horizontal) label: IMDb Rating
  - Y axis (vertical) label: Wins+Noms





Example for scatter plots, on a single HTML page.

Q1 Deliverables:
The directory structure should be organized as follows:

Q1/
scatterplot.(html / js / css)
explanation.txt
scatter_plots.pdf
movies.csv

- **scatterplot.(html / js / css)** - the html / js / css files created.
- **explanation.txt** - the text file explaining your observations for a.2 and c.

- **scatter_plots.pdf** - a PDF document showing the screenshots of the five scatter plots created above (two for a.1, one for b and two for c). You may print the HTML page as a PDF file, and each PDF page shows one plot (**hint:** use CSS page break). Clearly title the plots as instructed (see examples in figure).
- **movies.csv** - the dataset.