STAT 471: Programming Assignment 2

Due: October 4, 2025 at 11:59pm

1 Instructions

Please make sure to submit your solutions to the following questions in a knitted .html file format (or you can use an .nb.html if you'd like). Rmd files are acceptable as well.

2 Question 1 (50 points)

Your coworker wants to go see the Demon Slayer Infinity Castle movie and can't show up to work, so you cover his task for the day to perform data transformations on the IMDB movies dataset (imdb_top_1000.csv).

(a) Load in the dataset. Use the dplyr library's select() function to select the following column features: Series_Title, Released_Year, Runtime, Certificate, IMDB_Rating, and Meta_score.

(b) Filter your dataframe so that the Release_Year is after the year 1985.

(c) Use the transmute() function to create a new variable called "cert_multicat" and perform one-hot encoding for the following categories: PG, PG-13, TV-14, U, and R. Assign 1 to PG, 2 to PG-13, and so forth.

3 Question 2 (50 points)

Consider the airquality built-in dataset.

(a) Ensure that the Day and Month features are set as factors using factor(). For Days, make sure that the levels are from 1 to 31 using the c() function and set the ordering in the ordered parameter to be TRUE. For Months, the levels should be subsetted from May to September (5:9) and set the labels to be their respective month name using month.abb. Ordering should also be set to TRUE here.

(b) Create a histogram for the Ozone feature and compute the skewness of the Ozone distribution. Is our distribution left skewed, right skewed, or normal?

(c) Apply a square root transformation to the Ozone feature and replot the histogram. Recompute the skewness of Ozone. What happened after your transformation?

(d) Create a scatterplot between the transformed Ozone variable and the Wind features. Color the points by month. Is there a correlation between Ozone and Wind? If so, is it a positive or negative correlation? Use cor() to compute the correlation coefficient between the features to support your claim. Ensure that use = "complete.obs" to ignore missing values.