MCM ML 520 Assignment 1 - Part 2

Implementation Part 1 (50%): Diamond Prices

Dataset

Use the ggplot2 diamonds dataset (`diamonds.csv`). Original source: https://raw.githubusercontent.com/tidyverse/ggplot2/master/data-raw/diamonds.csv.

While it has a couple of thousand lines, this is a small dataset. If you are curious try loading, analyzing, and plotting it with a spreadsheet program, and extrapolate the lag that you experience to datasets with millions of lines.

Task

Make yourself familiar with the dataset by reading its minimal description at https://ggplot2.tidyverse.org/reference/diamonds.html.
Prepare an code script that fulfills the following tasks and answer the included questions. For each task put results into the report which support your answers (e.g. figures):

1. Give an overview of the dataset structure by answering those questions:
   - How many samples and features are in the dataset?
   - What are the feature data types?
   - Are diamonds balanced across color, cut and clarity? (Hint: roughly 1:1 means balanced, e.g. 1:2 is a "1:2 imbalance")
2. Visualize diamond prices using a histogram, boxplot and densityplot. Answer this question:
   - Is there trend visible in those plots? If yes, which is it and in which plots can you see it?
3. Calculate and state the mean, median, standard deviation, median absolute deviation (MAD), 1st and 3rd quartile (Q1 and Q3), and inner quartile range of the diamond price.
   - If you are not familiar with those functions: use Google, Wikipedia, etc.
   - Required commands are all in the provided script.
4. Plot the diamond price against the carat values as a scatterplot. Answer this question:
   - Is there a trend visible in the plot? If yes, which is it?

- Hint: plotting many samples will be slow. Changing the plot symbol to `'.'` will cause a speedup.

5. Analyze the correlation between diamond price and diamond x, y, and z dimensions. Answer those questions:
   - Create pairwise plots for these features.
   - Is there a trend visible between x, y, and z? If yes, which is it?
   - Is there a trend visible between the dimensions and the price? If yes, which is it?
   - Hint: if you don't know what a linear relation is (Google it!):
     - Linear correlation: feature A low –> feature B low, and feature A high –> feature B high.
     - (Inverse) linear correlation is also a linear correlation: feature A low –> feature B high, and feature A high –> feature B low: inverse linear correlation. Usually also just called linear correlation.
     - When plotting feature A against feature B and their points form a "straight line", then it's a linear relationship between A and B = linear correlation.
6. Analyze diamond prices per diamond color.
   - Create boxplots showing diamond price boxes for each diamond color (all boxes should be in one figure).
   - Create densityplots showing diamond prices for each diamond color (all densities should be in one figure).
   - Answer this question: is there a trend visible? If yes, which one?
7. Use vectorized commands (= *no loops!*) to answer these questions:
   - How many diamonds have a price above 9500?
   - How many diamonds have a price above 9500 and have color "D"?
   - What is the mean and std of the price of all color "D" diamonds with cut "Fair"?
   - What is the median and mad of the price of all color "J" diamonds with cut "Ideal"?
   - Create two copies of the dataframe that contains only the price and carat feature. Apply a log with base 10 to both features in one of those dataframes, and square ($x' = x^2$) the features in the other dataframe. What is the mean and std of the transformed features in both dataframes?

Implementation Part 2 (50%): Cell Body Segmentation Data

Dataset

Use the segmentationData dataset provided as csv-file. The class we differentiate is stated by the "`Class`" feature.

Task

Analyze the data using the same techniques as for the last task. Decide for yourself which and how to use the specific commands. Answer the following questions in the report and include figures supporting your answers:

1. Which classes exist? Are they (roughly) balanced?
2. Which noteworthy trends of features and relations between features as well as features and Class do you see?
3. If you would need to distinguish the classes with those features, which features would you choose, any why?