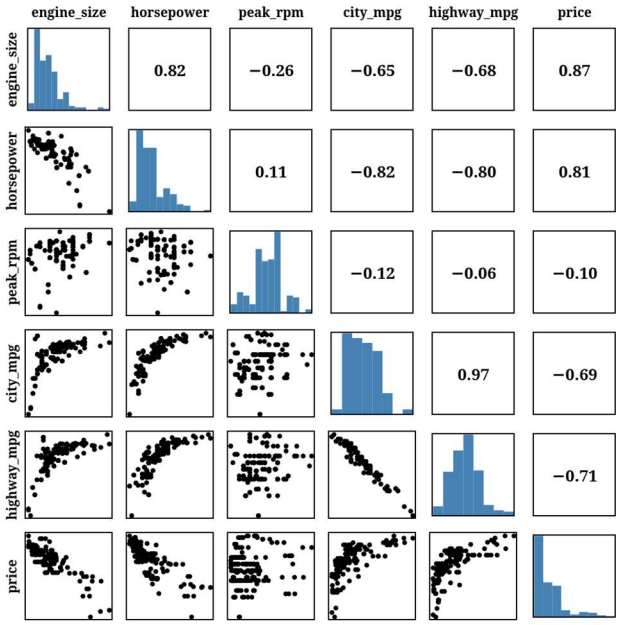
Exercise 9: Scatterplot Matrix **(20 points)**

Due: 10.07.2023 8AM

Contributor 1: Jonas Stettner
Contributor 2: Ana Sanchez Acosta



Task 1: Scatterplot Matrix (15 points)

Task 1a: Basics (10 points)

For this exercise, your task is to create a scatter plot matrix of the cars data set, as shown in the figure above. For comparisons of dimensions with themselves, add a histogram showing the distribution of values. For comparisons of dimensions with other dimensions, either create a scatter plot or calculate and show the sample Pearson correlation coefficient (see https://en.wikipedia.org/wiki/Pearson_correlation_coefficient# For_a_sample for details). If the row index is larger than the column index, show a scatter plot, if the column index is larger than the row index, calculate and show the correlation. Make sure to include axis labels. There is no template files given this time.
You may only use d3.js and no other additional libraries (for layout, correlation calculation, etc.). You may use your own previously written code as orientation.

Task 1b: Extension (5 points):

- Use a larger subset of the automobile data set. You can download the necessary files from https://archive.ics.uci.edu/ml/datasets/automobile. (1 point)
- Include the option to switch between the Pearson correlation coefficient and the Spearman's rank correlation coefficient (1 point)
- Include axis labels for the individual sub plots. Consider only showing the axis information on hovering to reduce clutter in the visualization.
- Make the dimensions sortable (1 point)
- Include highlighting on hovering over one of the visualization. That means, if you're hovering over the scatter plot of horsepower vs. engine_size, highlight the correlation coefficient of horsepower vs. engine_size as well as the histograms of horsepower and engine_size (1 point)

Task 2: Multivariate Data (5 points)

Imagine you are a data analyst working for a marketing research company. Your client, a leading e-commerce platform, has provided you with a large, multivariate dataset containing information about their customers. The dataset includes variables such as age, annual income, shopping frequency, and satisfaction rating. Your task is to explore the single values as well as pairwise relationships among the variables and construct targeted marketing strategies. You decide to create a visualization to gain insights into customer behavior. Also you want to justify your findings to your client, who is a non-expert on data visualization, by indicating the findings visually.

Which visualization would you choose to create?
Describe the visualization inlcuding visual encodings and justify your answer.

Answer:
Taking into account that we have 4 numerical dimensions, we would choose a scatterplot matrix that displays multiple scatterplots of pairs of variables. Using the position encoding in each scatterplot, one can explore if there are any pairwise relations between the give variables.

On the diagonal we would plot a histogram to explore the distribution of the individual values, where the frequency of a band of values is determined by position and length encoding.

Scatterplot matrices are easy to understand which would be suitable for the costumer. It also helps to identify two dimensional patterns and correlations as well as to detect outliers. We would use a small radius for each point to reduce visual clutter and since there are not a lot of variables it is still possible to have an overview of the data.

This visualization would make it possible to explore the data as well as presenting the findings to the costumer in a clear and organized manner.

Submission: Zipped folder including all necessary files to display the visualizations on one page