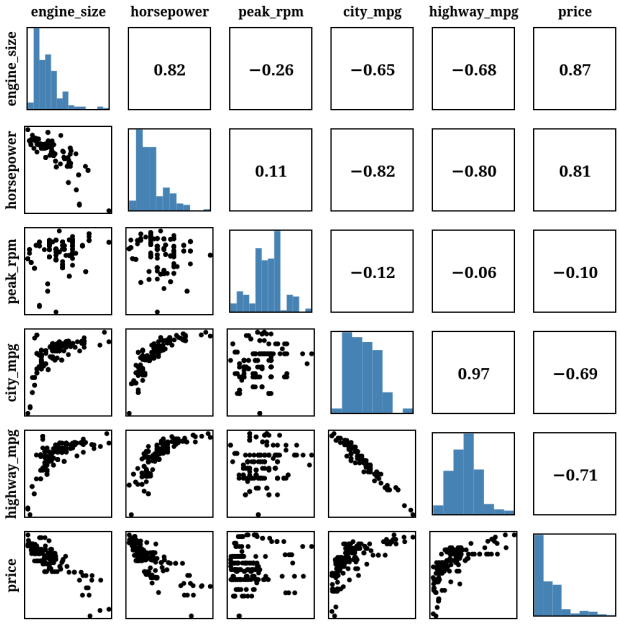
Exercise 6: Scatterplot Matrix **(20 points)**

Due: 10.06.2024 10AM

Contributor 1:
Contributor 2:



Task 1a: Basics (6 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, calculate and show the sample Pearson correlation coefficient (see https://en.wikipedia.org/wiki/Pearson_correlation_coefficient). 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 (4 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)
- 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 1c (5 points):
Using the visualization you created, list a number of findings about the datasets:

- Which ones are the correlated dimensions?
- What elements in the visualizations can help you identify the correlated dimensions?
- What can you say about the data, based on the correlations?

Task 2: Multivariate Data (5 points)

Imagine you are a data analyst working for a marketing research company. Your client, a fitness app developer, has provided you with a large, multivariate dataset containing information about their users. The dataset includes variables such as age, daily active minutes, weekly workout frequency, and user 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 user 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:

We would choose to create a scatterplot matrix for the visualization. Given that there are only four variables, the SPLOM would allow for the viewer to explore both pairwise correlations and single values through histograms in the diagonal.
One limitation of SPLOM it is difficult to display ordinal data (like weekly frequency or satisfaction), but this could be tackled by using jitter plots which would minimize overlapping and allow for better visualization of each data point.
This type of visualization would be best given that we have a large dataset (tabular view would be excesively large and therefore hard to read) and that it allows for the detection of all pairwise correlation (not needing reordering such as for PCP).

So the SPLOM would have a histogram with the data distribution on the diagonal, jigger plots in the inferior triangle and correlation analysis (like Pearson's) on the upper triangle. The latter would further summarize and add to the analysis to understand the relationships between variables.

When, for example, satisfaction rating does not have too many values, we could encode the data points with colors (in a sequential scale). In this way, the client could have a better insight of who are the people satisfied or unsatisfied, and create targeted recommendations. This could be further enhanced by interactive filtering.

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