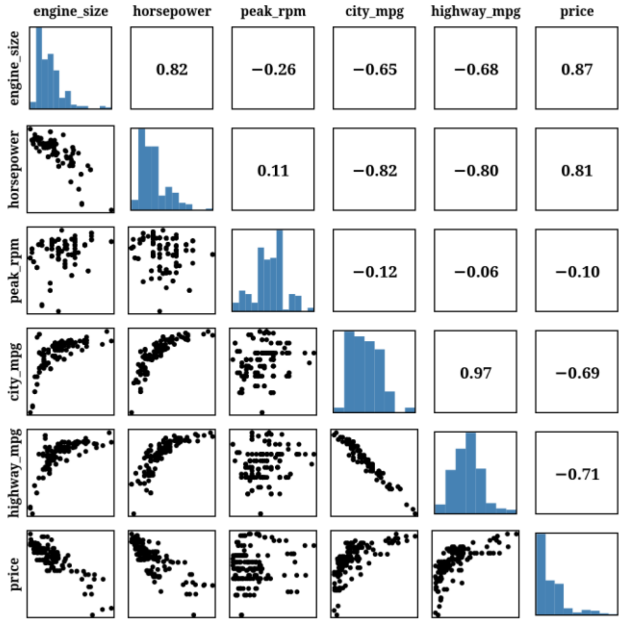
Exercise 9: Scatterplot Matrix (20 points)

Due: 10.07.2023 8AM

Contributor 1: Anurag Pacholi
Contributor 2: Mohammad Nomaan Husain



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:
We would choose the scatter plot matrix visualization. A scatter plot matrix allows us to explore both the single values and pairwise relationships among the variables simultaneously. It provides a clear and concise way to visualize the relationships between multiple variables in a single chart which makes it easier for a non-expert to understand what we would explain. With a large, multivariate dataset, it is crucial to understand the relationships between variables. The scatter plot matrix enables us to visualize these relationships in a comprehensive manner, providing a holistic view of customer behavior. By displaying scatter plots for each combination of variables, we can easily identify patterns, clusters, or trends in the data. This allows us to uncover potential correlations or associations that may exist among the variables, helping us identify important factors influencing customer behavior. The visual encodings of position and color make it easy to interpret the relationships between variables and identify any outliers or unusual patterns.

The visual encodings used to represent the data could include:
Position: The position of each data point on the x and y axes represents the values of the variables being compared. For example, if we have age on the x-axis and annual income on the y-axis, the position of each point will correspond to the specific age and income values for a customer.
Color: Color can be used to differentiate different variables or categories within the dataset. For example, we can assign different colors to represent different age groups or customer segments. By using color, we can visually identify and distinguish data points belonging to different groups.
Shape: The shape can be used as an encoding to represent different variables or categories. Different shapes, such as circles, triangles, or squares, can be assigned to different variables or groups, allowing us to visually differentiate and identify the data points.
Size: The size of the data points can also be used as an encoding to represent a certain variable or attribute. For instance, the size of each point could correspond to the shopping frequency or satisfaction rating of a customer. Larger points may indicate higher shopping frequency or satisfaction, while smaller points may indicate lower values.
Transparency/Opacity: Another encoding option is to use transparency or opacity to represent a variable or category. By adjusting the transparency of the data points, we can visualize overlapping points more effectively. This can be useful when there are dense clusters of data, and we want to show the concentration of points.

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