Homework: 5 Essential Graphs for Data Exploration

A Student from DataRockie Batch 11

2025-06-27

DataRockie Data Science Bootcamp Homework 

This document serves as my submission for the data visualization assignment for the **DataRockie Data Science Bootcamp, Batch 11**.

The goal of this exercise is to demonstrate proficiency in creating and interpreting fundamental graph types for exploratory data analysis using R and the `ggplot2` library. For this assignment, we will use the `mpg` dataset, which is included in `ggplot2`. This dataset contains fuel economy data for 38 popular models of cars.

First, we load the necessary libraries.

```
# We need these libraries for plotting and data manipulation.
library(ggplot2)
library(forcats) # Used for reordering the bar chart bars
```

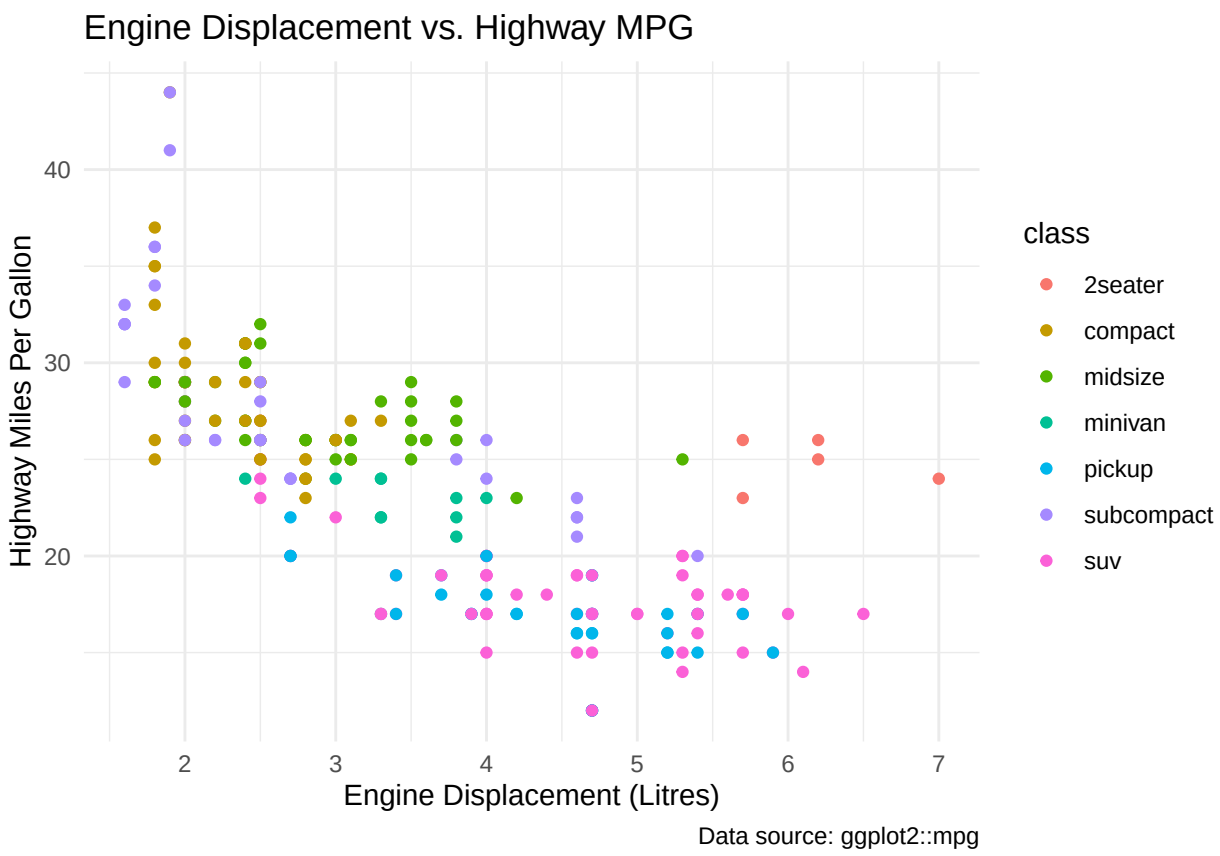
Graph 1: Scatter Plot

A scatter plot is used to visualize the relationship between two continuous (numeric) variables. Each point on the graph represents a single observation (in this case, a car).

Why is this graph useful? It helps us identify patterns such as correlation (positive or negative), clusters, and outliers. We can quickly see if an increase in one variable is associated with an increase or decrease in another.

Explanation of this graph: This scatter plot shows the relationship between a car's engine displacement (`displ`, in litres) and its highway fuel efficiency (`hwy`, in miles per gallon). We can clearly see a **negative correlation**: as the engine size increases, the highway miles per gallon tends to decrease. This makes intuitive sense—bigger engines are generally less fuel-efficient. I have added color to distinguish between the different car classes.

```
ggplot(data = mpg, aes(x = displ, y = hwy)) +
  geom_point(aes(color = class)) +
  labs(title = "Engine Displacement vs. Highway MPG",
       x = "Engine Displacement (Litres)",
       y = "Highway Miles Per Gallon",
       caption = "Data source: ggplot2::mpg") +
  theme_minimal()
```



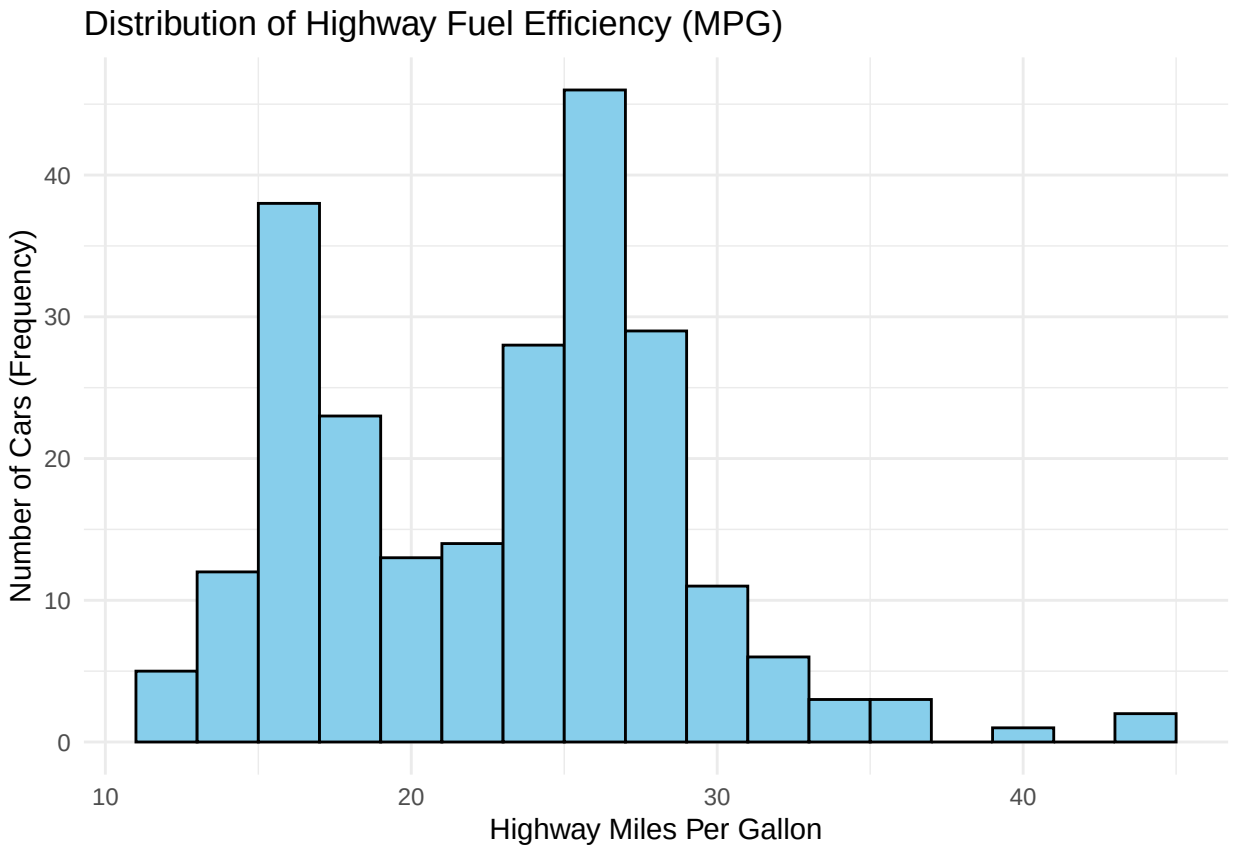
Graph 2: Histogram

A histogram is used to visualize the distribution of a single continuous variable. It groups numbers into ranges (called "bins") and the height of the bar shows the frequency (or count) of data points in that range.

Why is this graph useful? It allows us to understand the underlying frequency distribution of a variable. We can see where the majority of values are concentrated, the overall shape of the distribution (e.g., symmetric, skewed), and if there are any gaps or unusual values.

Explanation of this graph: This histogram shows the distribution of highway miles per gallon (`hwy`) for all cars in the dataset. The majority of cars have a fuel efficiency between 25 and 30 MPG. The distribution is **right-skewed**, meaning there is a long tail of cars with higher-than-average fuel efficiency, while most cars are clustered at the lower end.

```
ggplot(data = mpg, aes(x = hwy)) +
  geom_histogram(binwidth = 2, fill = "skyblue", color = "black") +
  labs(title = "Distribution of Highway Fuel Efficiency (MPG)",
       x = "Highway Miles Per Gallon",
       y = "Number of Cars (Frequency)") +
  theme_minimal()
```



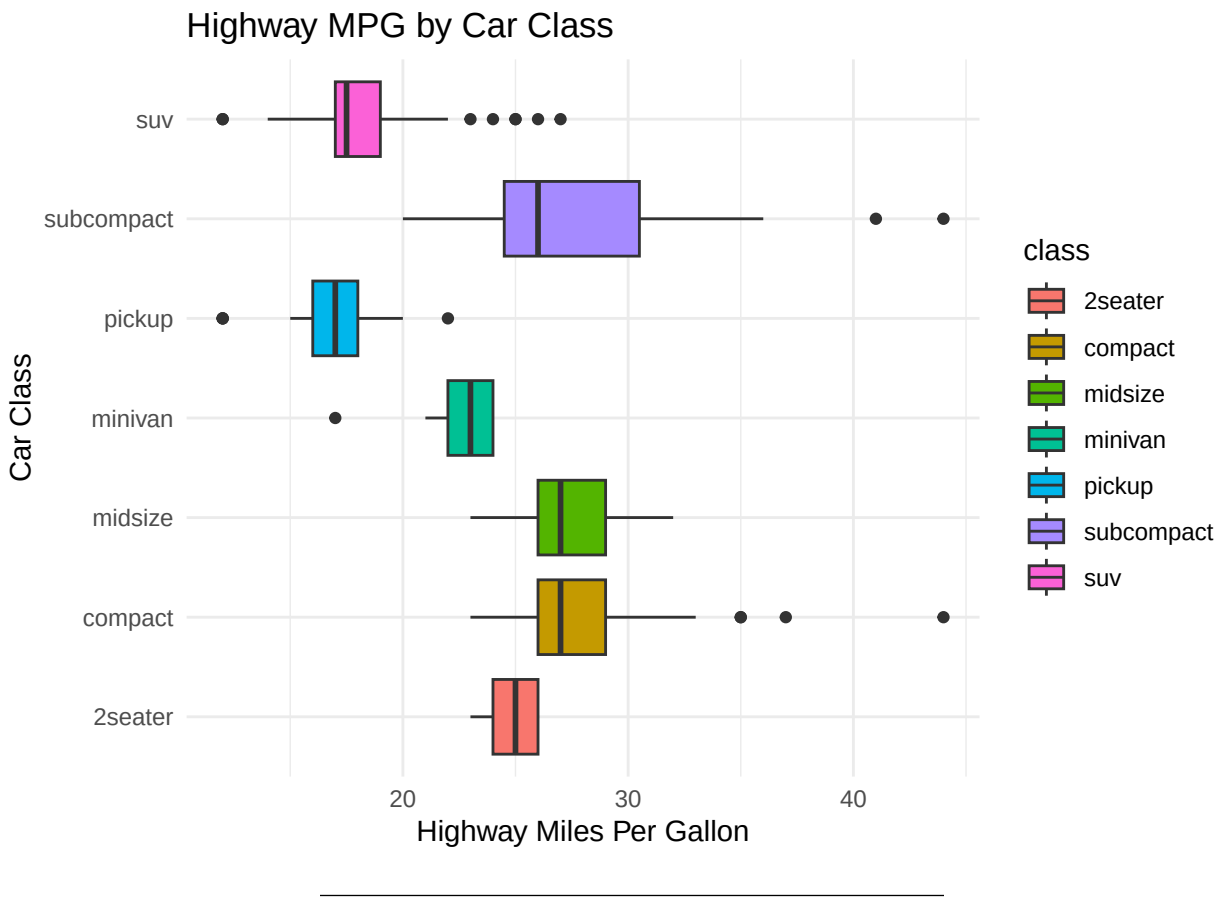
Graph 3: Box Plot

A box plot is excellent for comparing the distribution of a continuous variable across several categories of a categorical variable.

Why is this graph useful? It provides a concise summary of a distribution, showing the median (the center line), the interquartile range (the box, representing the middle 50% of data), and the overall range (the whiskers). It's particularly powerful for comparing groups.

Explanation of this graph: This box plot compares the highway fuel efficiency (`hwy`) across different car classes (`class`). It clearly shows that `compact` and `subcompact` cars tend to have the highest fuel efficiency, while `pickup` trucks and `suvs` have the lowest. The dots represent outliers—cars that have unusually high or low MPG for their class.

```
ggplot(data = mpg, aes(x = class, y = hwy, fill = class)) +
  geom_boxplot() +
  coord_flip() + # Flips coordinates to make class labels readable
  theme(legend.position = "none") +
  labs(title = "Highway MPG by Car Class",
       x = "Car Class",
       y = "Highway Miles Per Gallon") +
  theme_minimal()
```



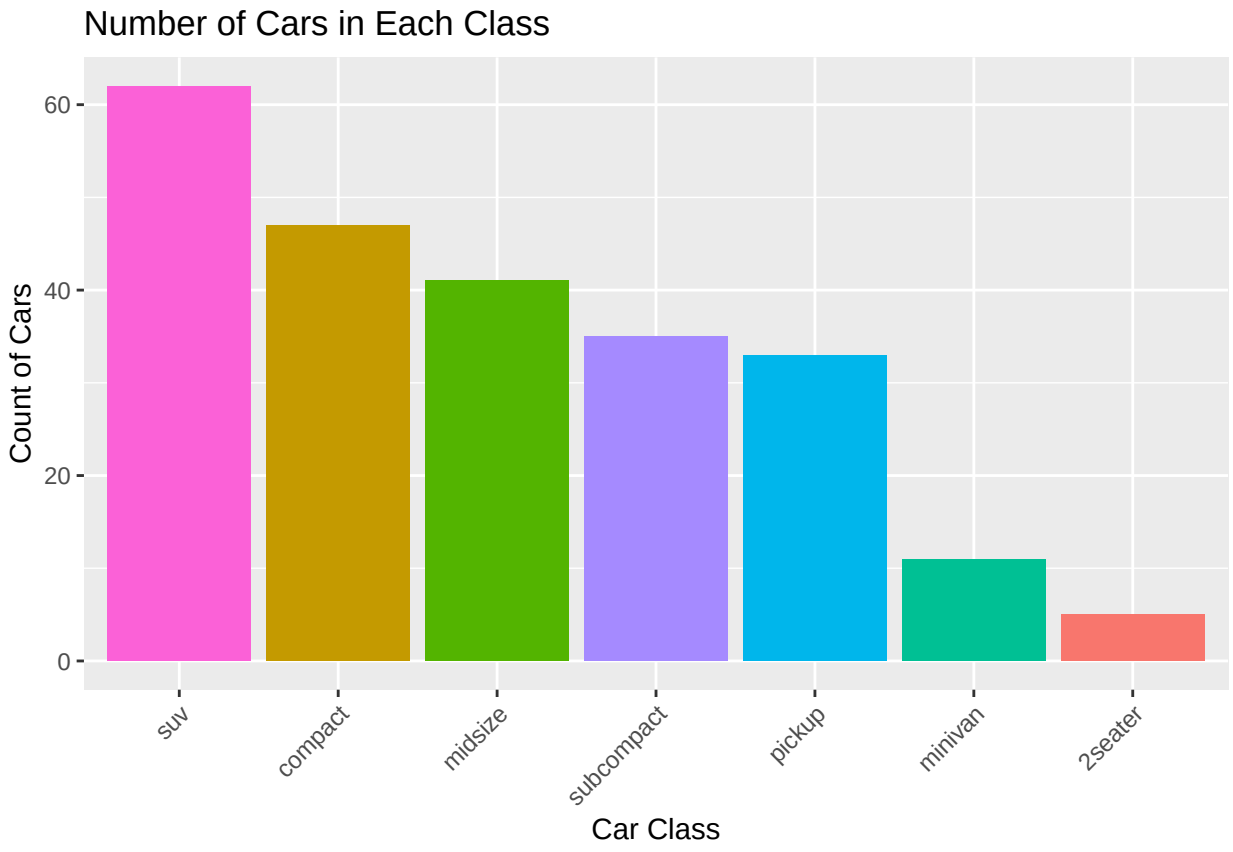
Graph 4: Bar Chart

A bar chart is used to display the frequency of a categorical variable. The length of each bar represents the count of observations within each category.

Why is this graph useful? It's a simple and effective way to compare the number of items in different groups. You can immediately see which categories are the most and least common.

Explanation of this graph: This bar chart shows the number of different car models for each `class` in the dataset. I have ordered the bars from most frequent to least frequent for easier comparison. The `SUV` class is the most represented category, while the `2seater` class is the least common.

```
ggplot(data = mpg, aes(x = fct_infreq(class), fill = class)) +
  geom_bar() +
  labs(title = "Number of Cars in Each Class",
       x = "Car Class",
       y = "Count of Cars") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        legend.position = "none")
```



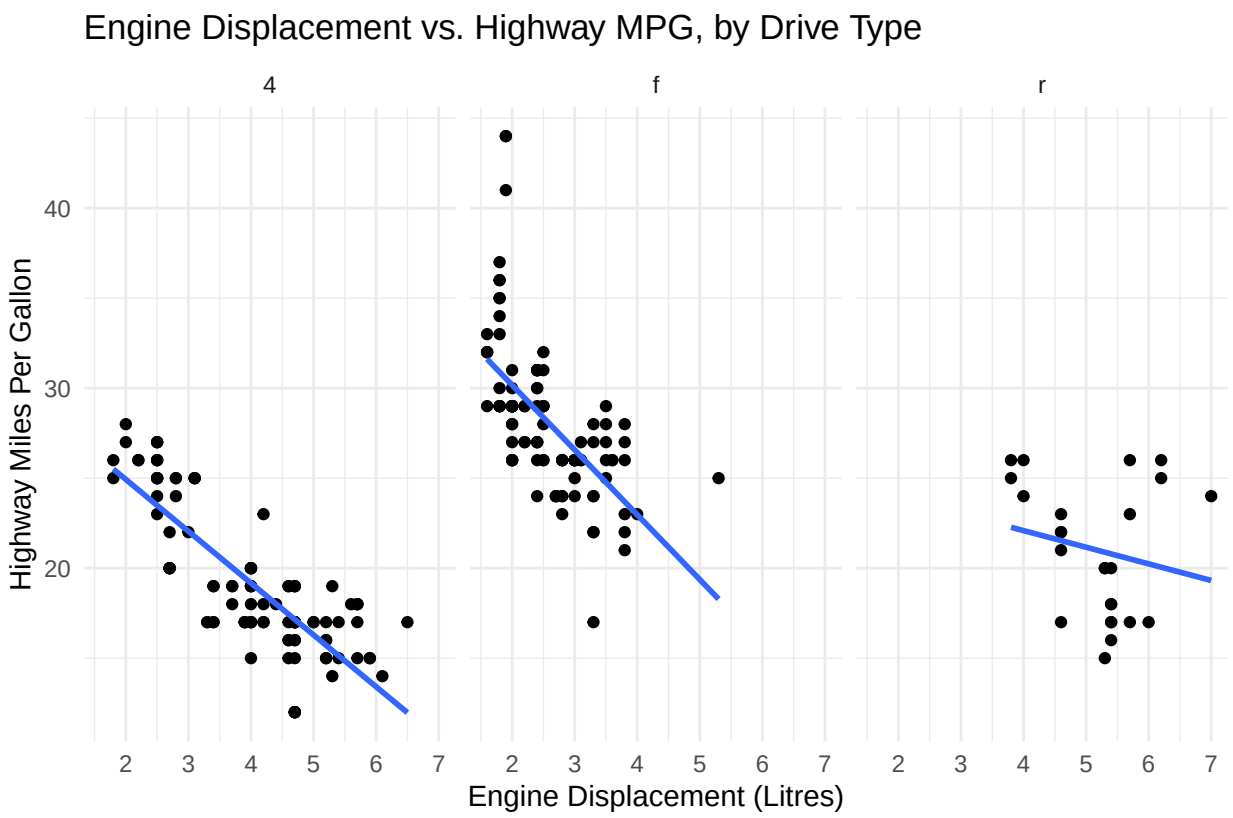
Graph 5: Faceted Plot

Faceting is a powerful technique where you break down one plot into multiple sub-plots. These sub-plots display subsets of the data, allowing for comparisons across a third variable.

Why is this graph useful? Faceting helps uncover more complex relationships. Instead of crowding one graph, we can see how a relationship (like in our scatter plot) changes depending on the category of another variable.

Explanation of this graph: This graph re-examines the relationship between engine displacement (`displ`) and highway MPG (`hwy`), but it is now **faceted** by the type of drive train (`drv`: 4-wheel, front-wheel, or rear-wheel). We can see that the negative correlation exists within all three drive types. However, it also reveals that for a given engine size, front-wheel drive (`f`) cars generally achieve higher MPG than rear-wheel (`r`) or 4-wheel drive (`4`) cars.

```
ggplot(data = mpg, aes(x = displ, y = hwy)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) + # Adds a linear model trend line
  facet_wrap(~ drv) +
  labs(title = "Engine Displacement vs. Highway MPG, by Drive Type",
       x = "Engine Displacement (Litres)",
       y = "Highway Miles Per Gallon",
       caption = "Each panel represents a different drive type.") +
  theme_minimal()
```


Engine Displacement vs. Highway MPG, by Drive Type
4
f
r
Highway Miles Per Gallon
40
30
20
2
3
4
5
6
7
Engine Displacement (Litres)
Each panel represents a different drive type.