Hussain Sarfraz

Dr. Kates

DATA-101-610-2021C

09/06/2021

Visualization Homework

To practice the skills we learned in this topic (Visualization), please solve the following problems:

1. Investigate the relationship between the number of cylinders (<cyl>) and highway fuel efficiency. Look at the variables, and decide which type of plot (scatterplot, line plot, boxplot, or bar chart) best summarizes their relationship. Comment on that relationship. HINT: you may need to use the as.factor(cyl) syntax in the graph (as we did for year above).

A boxplot would be the best graph to display the relationship between the number of cylinders(cyl) and highway fuel efficiency(hwy). The reason for this is because cyl is a categorical/discrete x-variable while hwy is a continuous y-variable.

To display a box plot I typed this in R:

NOTE: I had to add as.factor() around cyl to let R know that the cyl column type should be a factor/categorical variable and not a numeric type

INPUT

```
ggplot(data=mpg)+
  geom_boxplot(mapping=aes(x=as.factor(cyl),y=hwy))+coord_flip()
```

I then got this graph as a result:

OUTPUT

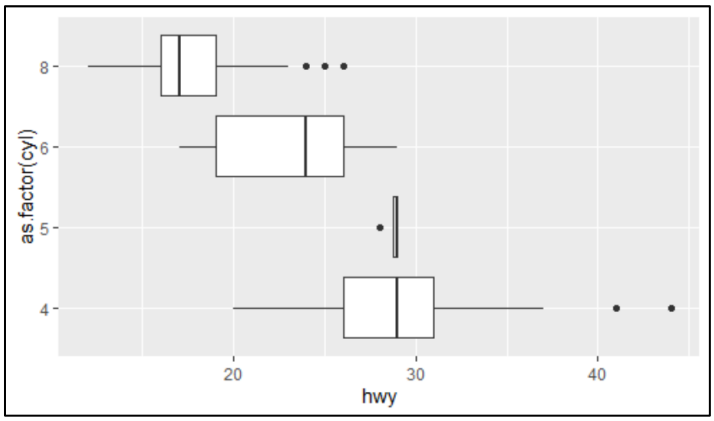

Figure 1: A boxplot that shows the distribution of fuel efficiency in each type of cylinder that a car has

From looking at the graph you can observe that cars with less cylinders have higher fuel efficiency since the median for the boxplot with 4 and 5 cylinders is approximately 28 miles per gallon. This result is way larger that the median milage shown for cars with 6 and 8 cylinders which is approximately 23 and 17 respectively.

I determined that cyl is a categorical/discrete variable because the number of cylinders can only be a whole number and can not include decimals. Also, when I entered the command "mpg$cyl" in R the number of cylinders shows these four repeated values only: 4,5,6,8. The number of cylinders is only limited to four groups based on the results of displaying all the values. For this reason, I have said that cyl is a categorical/discrete variable.

I determined that hwy is a continuous variable because highway miles (per gallon) can be presented as a decimal with an infinite value. I even typed the command "mpg$hwy" in R to see if my initial observations were correct and I saw that the values for highway fuel efficiency were not repeated like the values shown for the number of cylinders. There were many different values shown for hwy. These observations led me to say that hwy is a continuous variable

2. In section 3.9.1 of the textbook, solve problem #4 (on the relationship between city and highway fuel efficiency). What substantive conclusions can you draw about the relationship between these variables?

What does the plot below tell you about the relationship between city and highway mpg?

The plot tells us that the relationship between cty and hwy is strong. This means that no matter what type of vehicle you have the fuel efficiency will stay the same no matter which location you are in. In this plot you can see that each cars fuel efficiency was the same which is why the line increased as the scatter plot points increased. Also, the scatter plot points were close together

Why is `coord_fixed()` important?

coord_fixed() is important since it defines the aspect ratio of the line graph shown in problem #4. Adjusting a graphs aspect ratio helps with data visualization. The aspect ratio ensures that the x and y axis have a consistent ratio in a graph, no matter the size of the output window. In the example problem, the aspect ratio helped compress the data points and placed the line at a specific angle so observations and conclusions can easily be seen and made.

BEFORE coord_fixed()

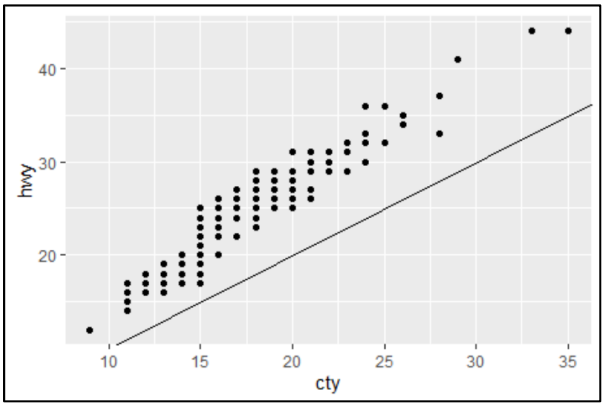

Figure 2: This is a graph without the coord_fixed() function. As you can see it is a bit hard to see any patterns since the data points are spread out a bit

AFTER coord_fixed()

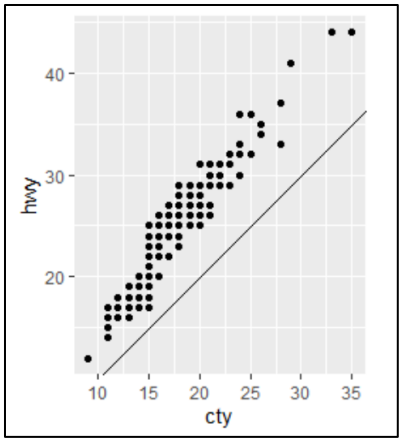

Figure 3: This is a graph with the coord_fixed() function. The points in this graph are closer together than in figure 2 which shows that using the coord_fixed() function allows an individual to make better observations about data since the patterns are easier to see. Also, the aspect ratio for the graph is set which would allow different users to see the graph in one form instead of in different sizes (a graphs initial size can be altered through a individuals output window size).

What does `geom_abline()` do?

geom_abline() creates a line which has a slope of 1. The line seems to correlate with the data in the scatter plot because when geom_abline() increases, the scatterplots increase as well. Because geom_abline() increases with the points in the scatter plot it can be concluded that a cars fuel efficiency does not change when in the highway or city. A cars fuel efficiency will be the same no matter where it is.

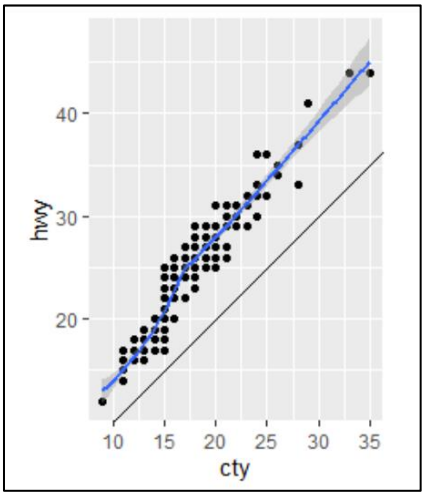

Figure 4: I added the geom_smooth() function to the graph to see if my observations were correct. The variables cty and hwy are correlated since they are closely surrounded by the blue line. The blue line is relatively straight and resembles the line created by the geom_abline() function.

3. Look at how the type of drivetrain influences fuel economy (<drv>). For a given engine size (<displ>), in general, do four-wheel drive, front wheel drive, or rear wheel drive engines have the highest fuel economy?

Front wheel drives have the highest fuel economy since the scatterplot in figure 5 shows that majority of the front wheel drivetrains are in the upper left-hand corner (green points in graph). Also, the front wheel drivetrains have a small engine size which reveals that cars with a small engine size have greater fuel efficiency.

There are some four-wheel drivetrains (red points in graph) that do mix with the green points, but majority of the red points are in the lower right-hand corner of the graph. This means that there are some four-wheel drivetrains that have a high fuel economy, but majority of the four-wheel drivetrains do not have a high fuel economy.

INPUT

```
ggplot(data=mpg)+
  geom_point(mapping=aes(x=displ,y=hwy,color=drv))
```

OUTPUT

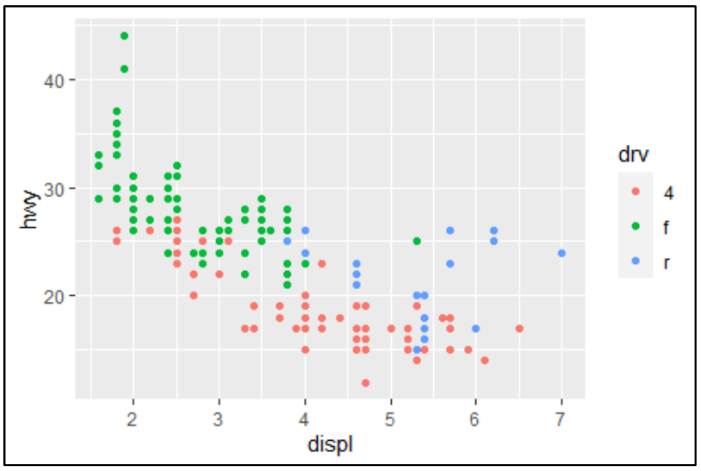

Figure 5: A scatterplot which compares the 2 continuous variables displ and hwy. The categorical variable drv color codes each point to show the type of drive that is represented by each point.

Feedback

1. You were exactly right here! Boxplot was the correct visualization to use here since it is helpful for us to see ranges in our continuous variable (<hwy>) for each value/group of our discrete variable (<cyl>), just like you said. You were also spot on with the assessment that as cylinders increase highway efficiency decreases. (3/3)
2. Exactly right on all your points here as well. The only additional thing we were looking for here was that highway efficiency and city efficiency, while positively and linearly related, are not always exactly the same. Since all our data points are just above the geom_abline() line with an intercept of 0, we know that highway mpg is always going to be about 5-10 mph more than city mpg. (2/3)
3. Awesome job here as well. Using a scatterplot here was perfect, and breaking out the drivetrain with your colors was just what we needed to assess the relationship. Keep up the good work! (4/4)