Exercise 1:

Q: Generate a scatterplot between these two variables. Does it capture the intuitive relationship you expected? What happens if you make a scatterplot of class vs drv? Why is the plot not useful?

Answer:

For plot 1, I plot a scatterplot using two variables which are displ and hwy. Since 'A car with low fuel efficiency consumes more fuel than a car with a high fuel efficiency when they travel the same distance', we can deduce that there will be negative relationship between displ and hwy. And the relationship shown in plot 1 is the same as what I expect.

When I plot class and drv in plot 2, it is hard to say the relationship between class and drv. For each drv of car, there is only one

The class and drv attributes are categorical. Therefore the plot shows the mapping between these categories.

Q: In the plot you made above, one group of points seems to fall outside of the linear trend. These cars have a higher mileage than you might expect. How can you explain these cars?

Answer:

These cars are beyond expectation and they have higher mileage than predicted.

These points are outliers. Outliers are the points that don't appear to fit, assuming that all the other points are valid. In this case, the new version of big engine is having higher efficiency of oil than old version. This might be one of the reasons these outliers appeared.

Exercise 1b:

Q: What conclusions can we make?

Answer:

For cars that are the same type, the distribution of engine size is not very discrete.

Q: Describe what the scale_x_log10() does. Why is it a more evenly distributed cloud of points now? (2-3 sentences.)

Answer:

Scale_*_continuous is the default scales for continuous x aesthetics. It is a kind of transformation working on x values and then change the scale of the data. For scale_x_log10(), it changes the scale of x axis with changing the value itself before statistics to log scale. Since the x

value have been stretched after changing the scale, the cloud of points will also be more evenly distributed.

Q: What does the dollar() call do? How can you find other ways of relabeling the scales when using scale_x_log10()?

Answer:

The dollar() formats numbers as currency and rounds values to dollars or cents.

Other ways of relabeling:

Scales : : comma_format();

Scales : : unit_format(unit, scale);

Scales : : percent_format()

Q: describe in your words what is going on

Answer:

The color of points is not yellow.

Q: Write down what all those arguments in geom_smooth(...) do?

Answer:

Geom_smooth is a smoothing line added as a layer onto the basic plot to help to see the trend of data. Gam smoothing is generalized additive mode smoothing working with a large number of points.

Q: What does fill = continent do? What do you think about the match of colors between lines and error bands?

Answer:

The fill of points will depend on the type of continent. The math of colors between lines and error bands will not happen.

Q: Notice how the above code leads to a single smooth line, not one per continent. Why?

Answer:

Because it adds color to geom_point and then add smooth line on to it.

Exercise 2:

Based on the graphs that I plotted in r, people that have jobs such as housemaid, management and retired and have high balance are more likely to purchase y; people that are single and married and have high balance are more likely to purchase y; people that education is primary and secondary and have high balance are more likely to purchase y; people that do not have credits in default and have high balance are more likely to purchase y.