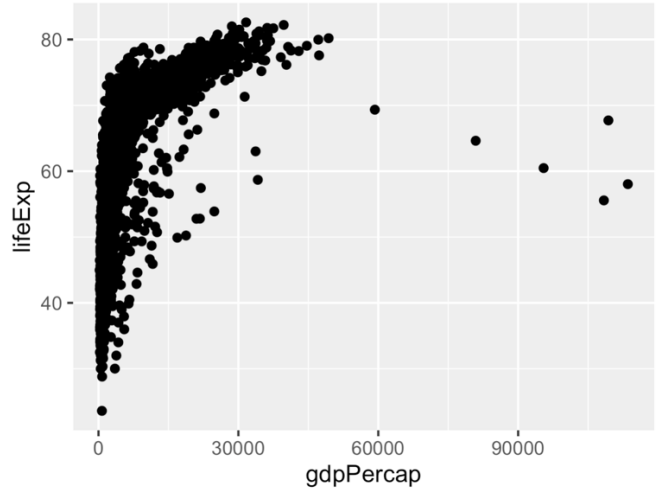
SSC 442 / Lab 1

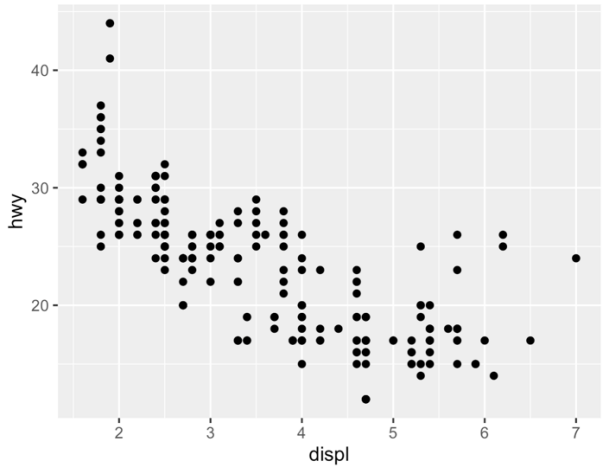
Shun Hu, Yuhan Liu, Yanfei Du, Hongji Li



Try it: What other types of plots are there? Try to find several more geom_

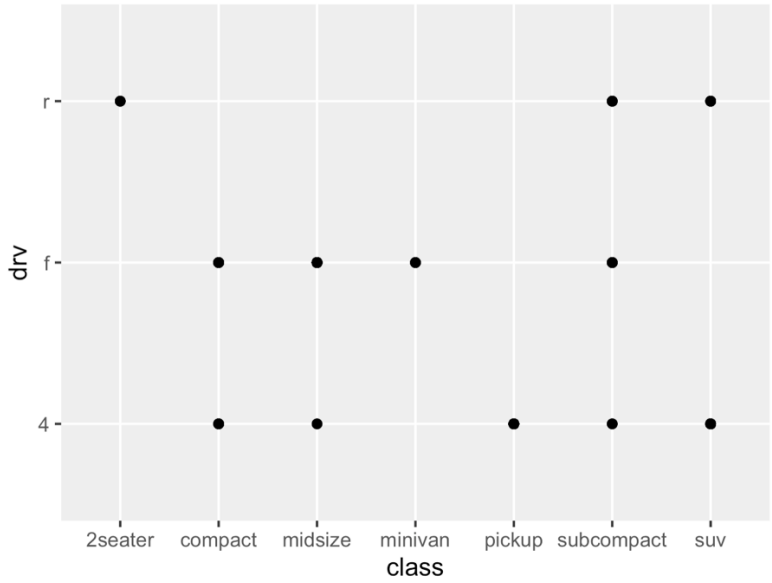
Answer: Line: geom_line() Histogram: geom_histogram()

Exercise 1



Does it capture the intuitive relationship you expected?

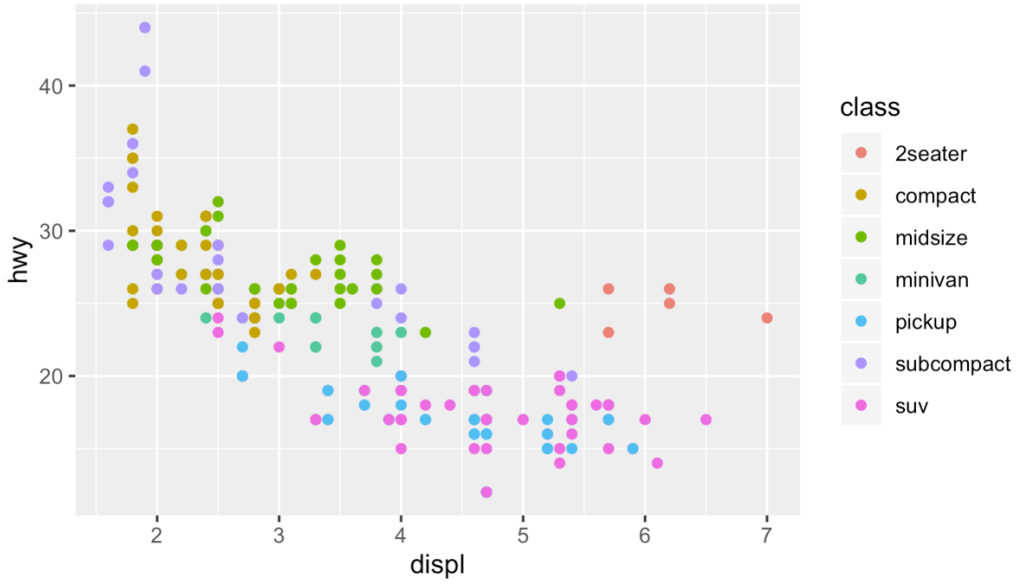
Yes, as a car's engine size increases, a car’s fuel efficiency on the highway decreases.



What happens if you make a scatterplot of class vs drv ? Why is the plot not useful?

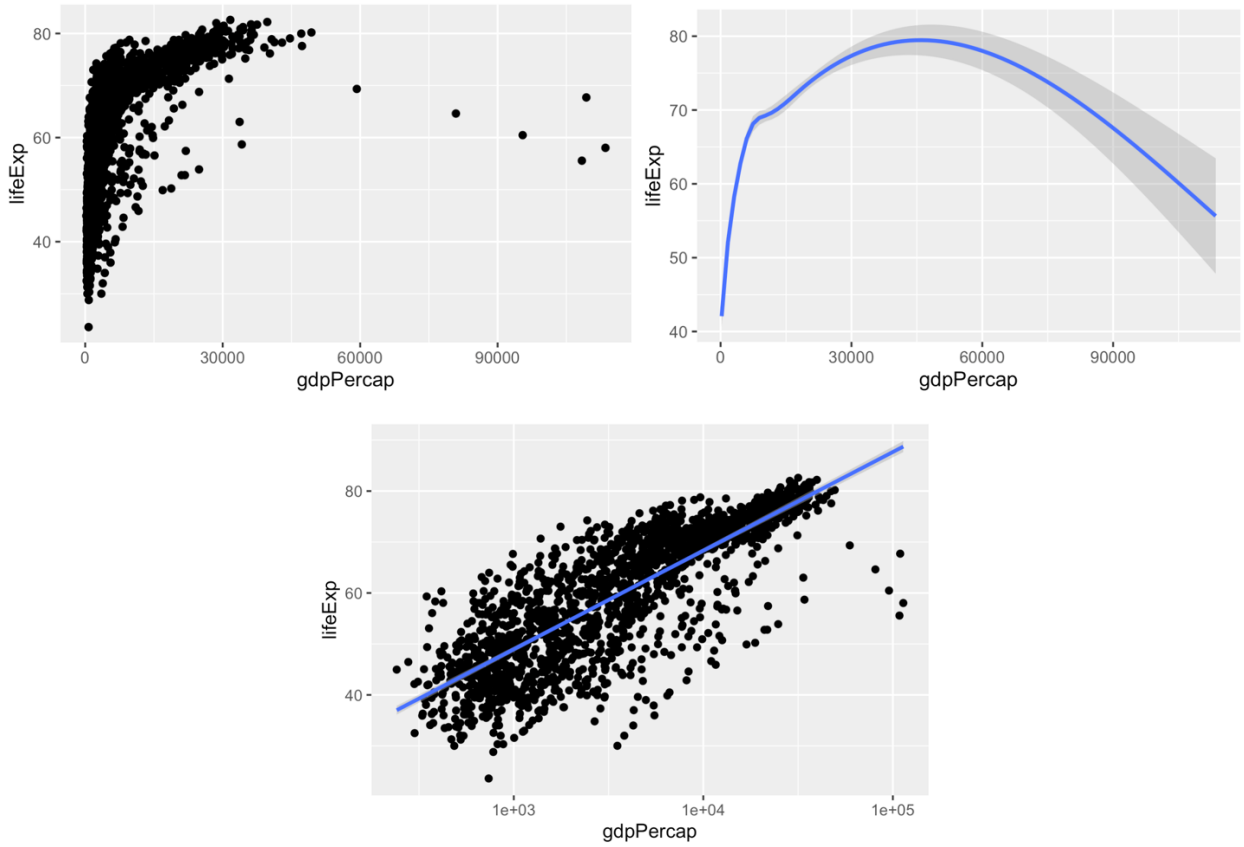
There is no linear trend between class and drv.

Exercise 1b



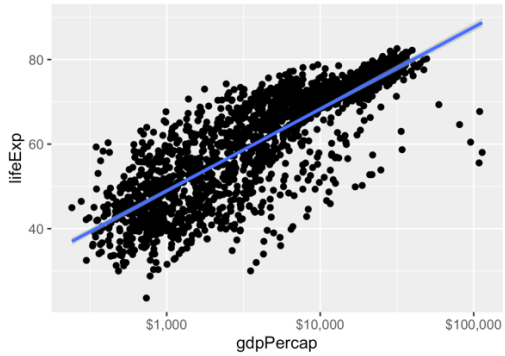
What conclusions can we make?

Smaller cars tend to have high hwy and low displ, while bigger cars tend to have low hwy and high displ.



Try it: Describe what the scale_x_log10() does. Why is it a more evenly distributed cloud of points now?

scale_x_log10() makes the original x values become log10(x), which redistribute data on x-axis more evenly. Then the picture looks more clear and concise.



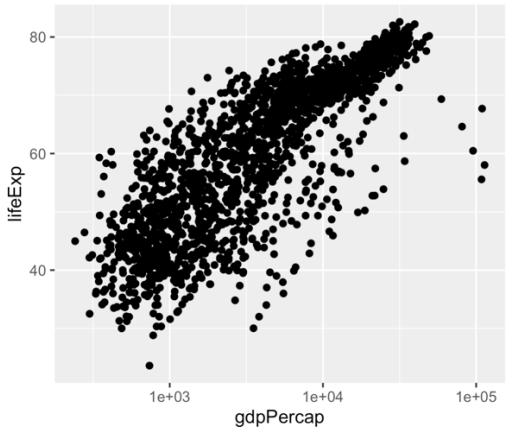
Try it: What does the dollar() call do?

The dollar() changes the labels of x-axis to dollors.

How can you find other ways of relabeling the scales when using scale_x_log10() ?

Change the value after scales:: , to number for instance.

Try it: Write the above sentence for the original call aes(x = gdpPercap, y = lifeExp, color = 'yellow')

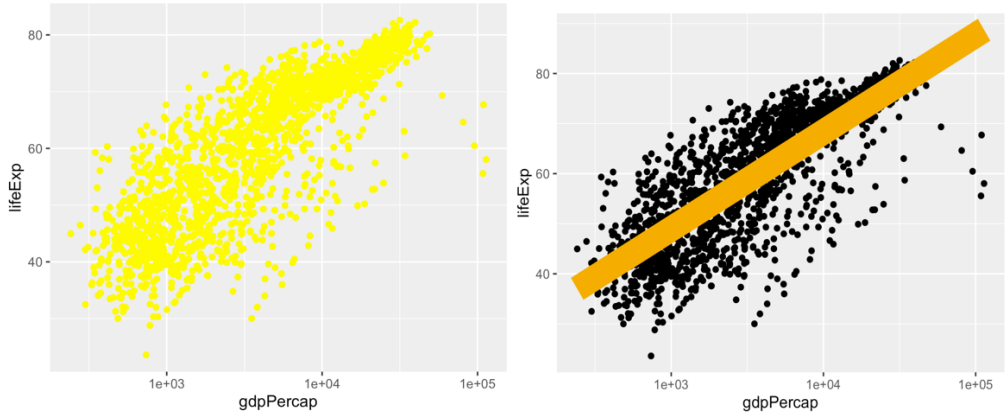


Try it: describe in your words what is going on.

The points in the graph is not yellow.

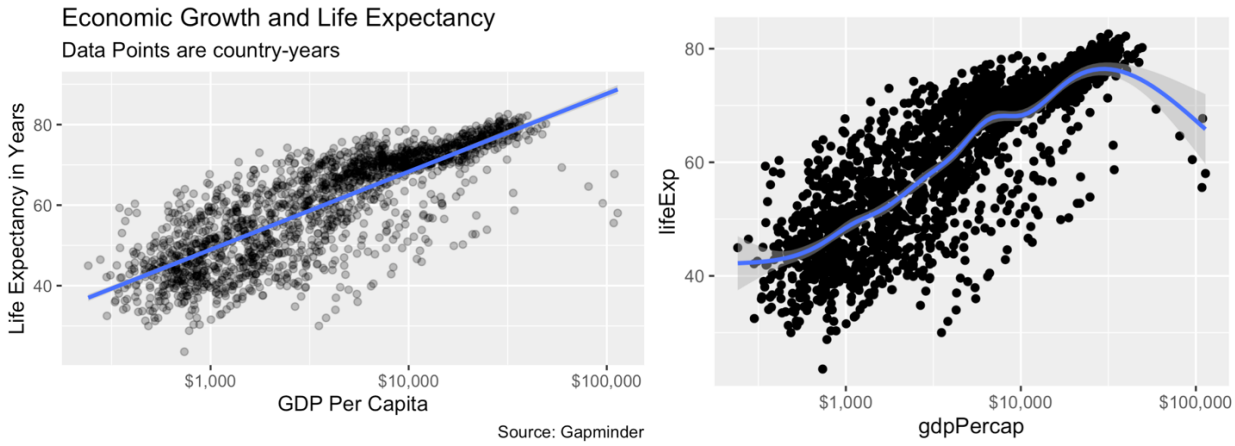
How can we tell ggplot to draw yellow points?

p + geom_point(color="yellow") + scale_x_log10()



Try it: Write down what all those arguments in geom_smooth(...) do

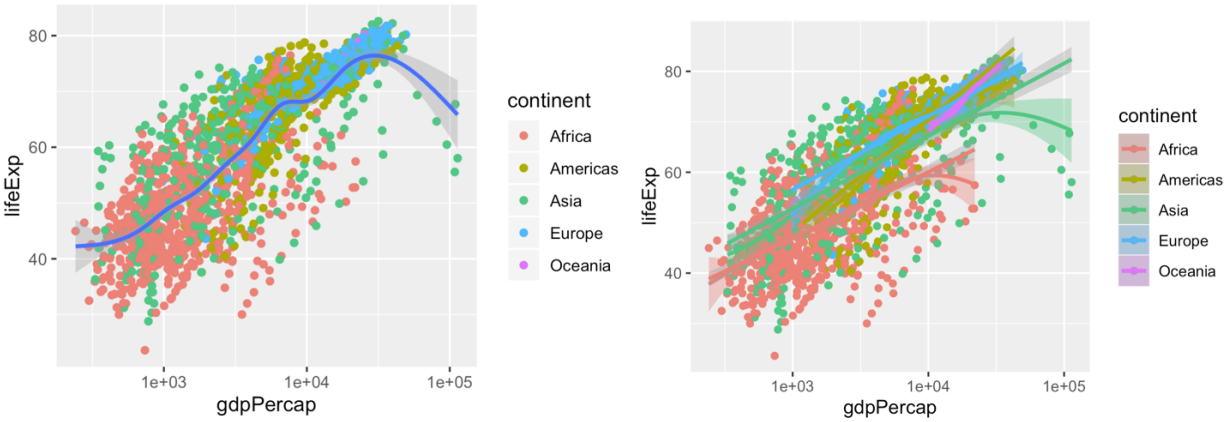
geom_smooth(...) generates a linear regression line for the given data and plot the line. Display confidence interval around smooth? color: color of the regression line. method="lm" : a function of smooth called lm.



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

Color the error bands of the regression line.

Match of colors between lines and error bands are important, in this way people can easily distinguish different lines and its error bands.



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

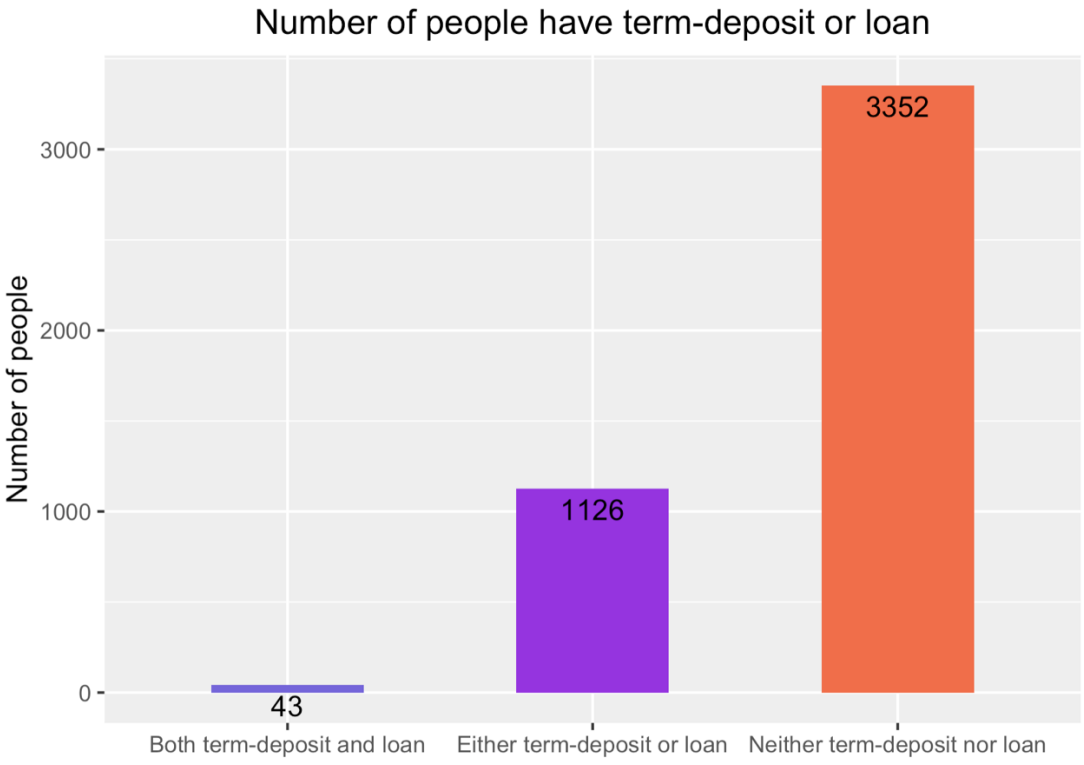
Because different continents' data are in the same dataframe, and we do not separate them when making the smooth line.

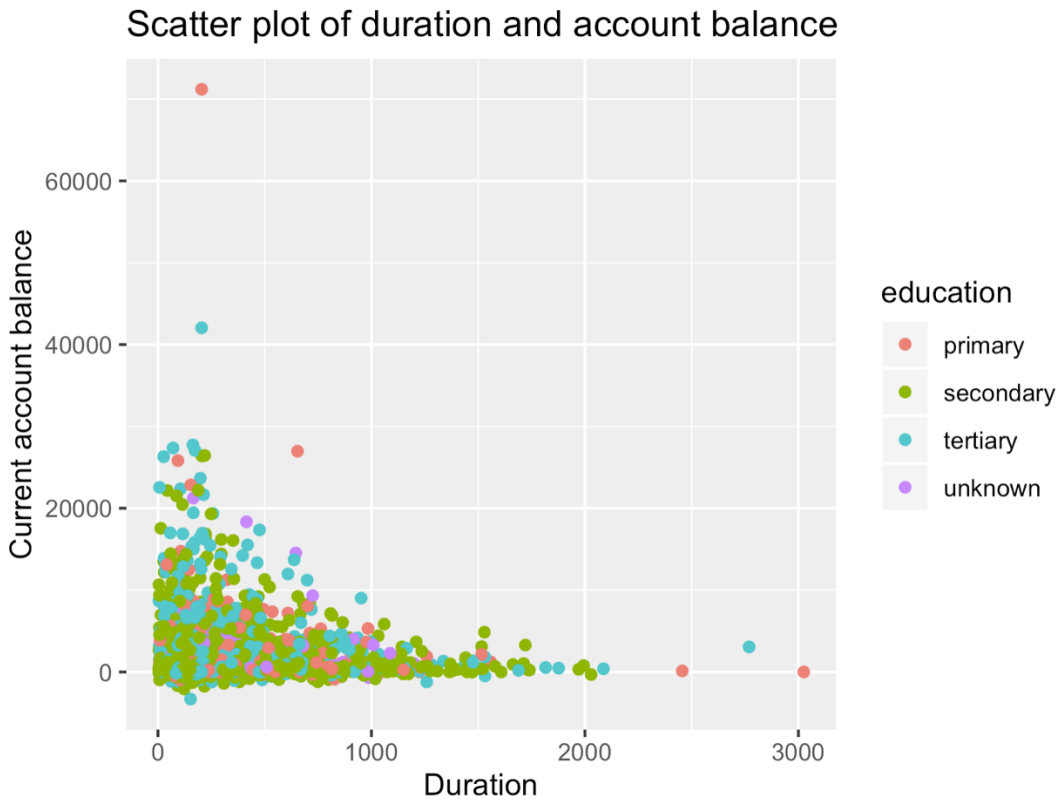
Try it: What is bad about the following example, assuming the graph is the one we want?

The above codes make regression line for each continent, and this makes the picture look very confusing. So we should set aesthetics at the top level rather than at the individual geometry level.

Setting aesthetics at the individual geometry level instead of top level make the code very confusing, and you have to set aesthetics at every single line which increase a lot of works.

Ecercise 2


Number of people have term-deposit or loan
Number of people
3000
2000
1000
0
3352
1126
43
Both term-deposit and loan
Either term-deposit or loan
Neither term-deposit nor loan


Scatter plot of duration and account balance
Current account balance
60000
40000
20000
0
0
1000
2000
3000
Duration
education
primary
secondary
tertiary
unknown

Analytics Memo

Purpose of analytics:

We already have the dataset of bank from the professor, thus we omit the part of collecting data. To help the bank increase their long term deposits, we need to know how many of our existing customers have term deposits. At the same time, we also want to know how many current customer have loans. Then we need to find out what variables are related to customers' deposit amount, such as education level, account opening duration, etc.

What can we tell from the two graphs?

The first graph shows that only very few customers have both loans and long-term deposits, and the majority of customers have neither loans nor long-term deposits. This shows that we still have great potential in existing customers. The second graph shows the relationship between the customer's deposit amount and the duration of account opening time. As we can see from the figure, the longer the account is opened, the less the deposit amount. This may be because our bank's service is not good, or our long-term deposit rate is relatively low. So existing customers have transferred money to other banks. At the same time, in the second graph, we can find that the higher the education level of customers, the more deposits amount they have.

Suggestions:

1. Increase long-term deposit interest rate
2. Improve service quality to keep existing customers
3. Target future customers to those people with higher education
4. Make some advertisements to stimulate existing customers to make long-term deposits