SSC442 Project Lab 1

Name: Jianqi Liu

Dylan Patel

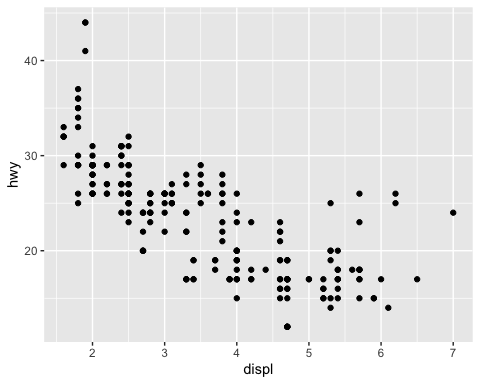
Ave’r Mckay

**Exercise 1:** Let’s briefly return to the mpg data. Among the variables in mpg are:

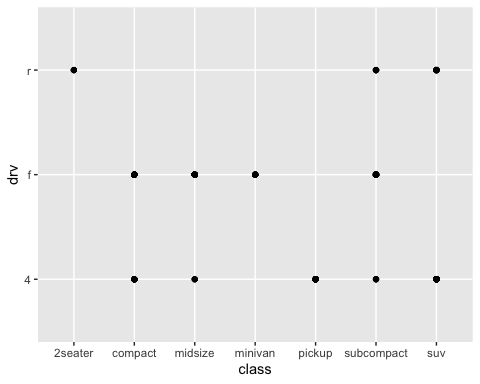
* displ, a car’s engine size, in litres.
* hwy, a car’s fuel efficiency on the highway, in miles per gallon (mpg). A car with a low fuel efficiency consumes more fuel than a car with a high fuel efficiency when they travel the same distance.

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?

library(tidyverse)  
ggplot(data=mpg,  
 mapping=aes(x=displ, y=hwy)) +  
 geom\_point()



# Yes, it capture the intuitive relationship I expected.  
# Bigger engine leads to higher consumption of gas  
  
ggplot(data=mpg,  
 mapping=aes(x=class, y=drv)) +  
 geom\_point()



# This scatterplot has no obvious meaning, because ‘drv’ and ‘class’ are discrete variables and can only take a few values, so it is not suitable to use scatterplot to mine information. You can consider using barplot

**Exercise 1b:** Using your previous scatterplot of displ and hwy, map the colors of your points to the class variable to reveal the class of each car. What conclusions can we make?

ggplot(data=mpg,  
 mapping=aes(x=displ, y=hwy)) +  
 geom\_point(aes(color=class))



# Conclusion: There are obvious differences between displ and hwy for different classes of cars:

# The most fuel-efficient models are subcompact and compact, and the most fuel-efficient models are suv and pickup

# 2seater's car has the largest displ on average, but the fuel consumption is not the highest (hwy is in the middle position), probably because this car is smaller and lighter

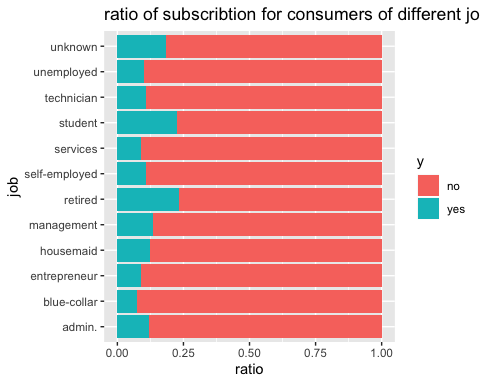
**Exercise 2**: Imagine that you have been hired as a consultant for a Portuguese bank. This bank recently undertook a series of marketing campaigns designed to increase deposits in their long-term savings accounts. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to assess if the consumer subscribed to the product (bank term deposit; coded as yes and no in the column y). As a result of this out-reach, we have the dataset bank.csv, which you can find here. I have given you a subset of a larger dataset. Most of the data is self-explanatory. However, a few inputs require that I give you more information: default describes if the customer has credit in default. housing and loan describe if the customer has either a mortgage or another type of loan already with the bank.

You must read this data into R and provide two data visualizations that are pertinent to making business decisions on this dataset. Describe, in a one-page memo, what information you hope to convey to executives about the outreach or about the dataset more generally.

bank <- read\_csv("bank.csv")

**visualization 1**

ggplot(data=bank,  
 mapping=aes(x=job, fill=y)) +  
 geom\_bar(position="fill") +  
 labs(title="ratio of subscribtion for consumers of different jobs",  
 y="ratio") +  
 coord\_flip()



# The figure above shows the distribution of the population of different jobs on the value of y

# It can be seen that the three categories of people "retired" and "student" and "unknown" are the most likely to accept the product,

# Of which "retired" is the highest, "student" is the second highest, and "unknown" is the third highest

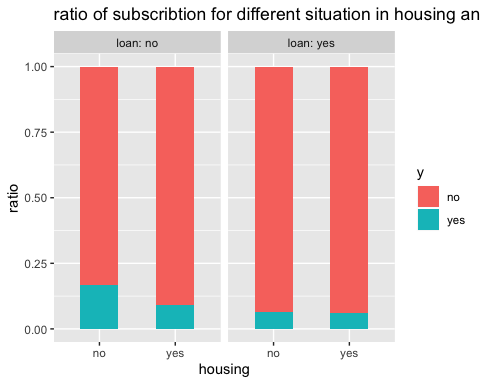
# The reason for this may be: Retired people generally have more spare money, and students may be unsteady and easily sold.

# Acceptance rates for the remaining groups are significantly lower than these three groups

# The lowest acceptance is "blue-collar", probably due to less free money

**visualization 2**

loanLabel = labeller(loan=c("no"="loan: no","yes"="loan: yes"))  
ggplot(data=bank,  
 mapping=aes(x=housing, fill=y)) +  
 geom\_bar(position="fill", width=0.5) +  
 facet\_grid(.~loan, labeller=loanLabel) +  
 labs(title="ratio of subscribtion for different situation in housing and loan",  
 y="ratio")



# The chart above shows the acceptance rate of bank products for people with and without their own homes, and those with and without loans.

# As can be seen from the figure:

# People with existing loans have a significantly lower acceptance rate than people without loans. This may be because the people with the loans have less spare money in their accounts.

# For people without loans, the acceptance rate of people who already have a house is lower than that of people who do not have a house, probably because they have less spare money after buying a house

# You can also see:

# Among those without loans, the effect of having a house on the acceptance rate is more obvious

# Among the people who carry loans, whether there is a house has no significant effect on the acceptance rate, and the ratio is almost the same