```
# ------------------------------------------------------------
#
# Anscombe's Quartet/Stratification
#
# ------------------------------------------------------------

library(tidyverse)
library(dslabs)
library(dplyr)
library(ggplot2)
library(Lahman)
library(HistData)

# Correlation is not always a good summary of the relationship
# between two variables. A famous example used to illustrate this are
# the following for artificial data sets, referred to as Anscombe's
# quartet. All of these pairs have a correlation of 0.82.

# Correlation is only meaningful in a particular context.
# To help us understand when it is that correlation is meaningful
# as a summary statistic, we'll try to predict the son's height using
# the father's height. This will help motivate and define linear
# regression. We start by demonstrating how correlation can be useful
# for prediction. Suppose we are asked to guess the height of a randomly
# selected son.

# Because of the distribution of the son height is approximately normal,
# we know that the average height of 70.5 inches is a value with the
# highest proportion and would be the prediction with the chances of
# minimizing the error.

# But what if we are told that the father is 72 inches?   Do we still
# guess 70.5 inches for the son?  The father is taller than average,
# specifically he is 1.14 standard deviations taller than the average
# father. So shall we predict that the son is also 1.14 standard
# deviations taller than the average son?   It turns out that this would
# be an overestimate.

# To see this, we look at all the sons with fathers who are about 72
# inches. We do this by stratifying the father's side. We call this a
# conditional average, since we are computing the average son height
# conditioned on the father being 72 inches tall. A challenge when using
# this approach in practice is that we don't have many fathers that are
# exactly 72. In our data set, we only have eight.

# If we change the number to 72.5, we would only have one father who is
# that height. This would result in averages with large standard errors,
# and they won't be useful for prediction for this reason.

# But for now, what we'll do is we'll take an approach of creating strata
# of fathers with very similar heights. Specifically, we will round
# fathers' heights to the nearest inch. This gives us the following
# prediction for the son of a father that is approximately 72 inches tall.
# We can use this code and get our answer, which is 71.84.

# This is 0.54 standard deviations larger than the average son, a smaller number
# than the 1.14 standard deviations taller that the father was above the average
# father. Stratification followed by box plots lets us see the distribution of
# each group. Here is that plot.

conditional_avg <- galton_heights %>% filter(round(father) == 72) %>%
  summarize(avg = mean(son)) %>% .$avg
conditional_avg
# [1] 71.83571

galton_heights %>% mutate(father_strata = factor(round(father))) %>%
  ggplot(aes(father_strata, son)) +
  geom_boxplot() +
  geom_point()

# We can see that the centers of these groups are increasing with height, not
# surprisingly. The means of each group appear to follow a linear relationship.
# We can make that plot like this, with this code.

galton_heights %>%
  mutate(father = round(father)) %>%
  group_by(father) %>%
  summarize(son_conditional_avg = mean(son)) %>%
  ggplot(aes(father, son_conditional_avg)) +
  geom_point()
```

```
# See the plot and notice that this appears to follow a line. The slope of this
# line appears to be about 0.5, which happens to be the correlation between
# father and son heights.

r <- galton_heights %>% summarize(r = cor(father, son)) %>% .$r
galton_heights %>%
  mutate(father = round(father)) %>%
  group_by(father) %>%
  summarize(son = mean(son)) %>%
  mutate(z_father = scale(father), z_son = scale(son)) %>%
  ggplot(aes(z_father, z_son)) +
  geom_point() +
  geom_abline(intercept = 0, slope = r)

# This is not a coincidence. To see this connection, let's plot the standardized
# heights against each other, son versus father, with a line that has a slope
# equal to the correlation. Here's the code. Here's a plot.

r <- galton_heights %>% summarize(r = cor(father, son)) %>% .$r
galton_heights %>%
  mutate(father = round(father)) %>%
  group_by(father) %>%
  summarize(son = mean(son)) %>%
  mutate(z_father = scale(father), z_son = scale(son)) %>%
  ggplot(aes(z_father, z_son)) +
  geom_point() +
  geom_abline(intercept = 0, slope = r)

# This line is what we call the regression line. In a later video, we will
# describe Galton's theoretical justification for using this line to estimate
# conditional means. Here, we define it and compute it for the data at hand.

# The regression line for two variables, x and y, tells us that for every
# standard deviation sigma x increase above the average mu x. For x, y grows rho
# standard deviations sigma y above the average mu y.

# The formula for the regression line is therefore this one. If there's perfect
# correlation, we predict an increase that is the same number of SDs. If there's
# zero correlation, then we don't use x at all for the prediction of y. For
# values between 0 and 1, the prediction is somewhere in between. If the
# correlation is negative, we predict a reduction, instead of an increase.

# It is because when the correlation is positive but lower than the one, that we
# predict something closer to the mean, that we call this regression. The son
# regresses to the average height. In fact, the title of Galton's paper was
# "Regression Towards Mediocrity in Hereditary Stature."

# Note that if we write this in the standard form of a line, y equals b plus mx,
# where b is the intercept and m is the slope, the regression line has slope rho
# times sigma y, divided by sigma x, and intercept mu y, minus mu x, times the
# slope.

# So if we standardize the variable so they have average 0 and standard
# deviation 1. Then the regression line has intercept 0 and slope equal to the
# correlation rho. Let's look at the original data, father son data, and add the
# regression line. We can compute the intercept and the slope using the formulas
# we just derived.

mu_x <- mean(galton_heights$father)
mu_y <- mean(galton_heights$son)
s_x <- sd(galton_heights$father)
s_y <- sd(galton_heights$son)
r < cor(galton_heights$father, galton_heights$son)
m <- r * s_y / s_x
b <- mu_y - m * mu_x

# Here's a code to make the plot with the regression line.

galton_heights %>%
  ggplot(aes(father, son)) +
  geom_point(alpha = 0.5) +
  geom_abline(intercept = b, slope = m)

# If we plot the data in standard units, then, as we discussed, the regression
# line as intercept 0 and slope rho. Here's the code to make that plot.

galton_heights %>%
  ggplot(aes(scale(father), scale(son))) +
  geom_point(alpha = 0.5) +
  geom_abline(intercept = 0, slope = r)

# We started this discussion by saying that we wanted to use the conditional
# means to predict the heights of the sons. But then we realized that there were
# very few data points in each strata. When we did this approximation of
# rounding off the height of the fathers, we found that these conditional means
# appear to follow a line. And we ended up with the regression line. So the
# regression line gives us the prediction.

# An advantage of using the regression line is that we used all the data to
# estimate just two parameters, the slope and the intercept. This makes it much
# more stable. When we do conditional means, we had fewer data points, which
# made the estimates have a large standard error, and therefore be unstable. So
# this is going to give us a much more stable prediction using the regression
# line. However, are we justified in using the regression line to predict?
# Galton gives us the answer.
```