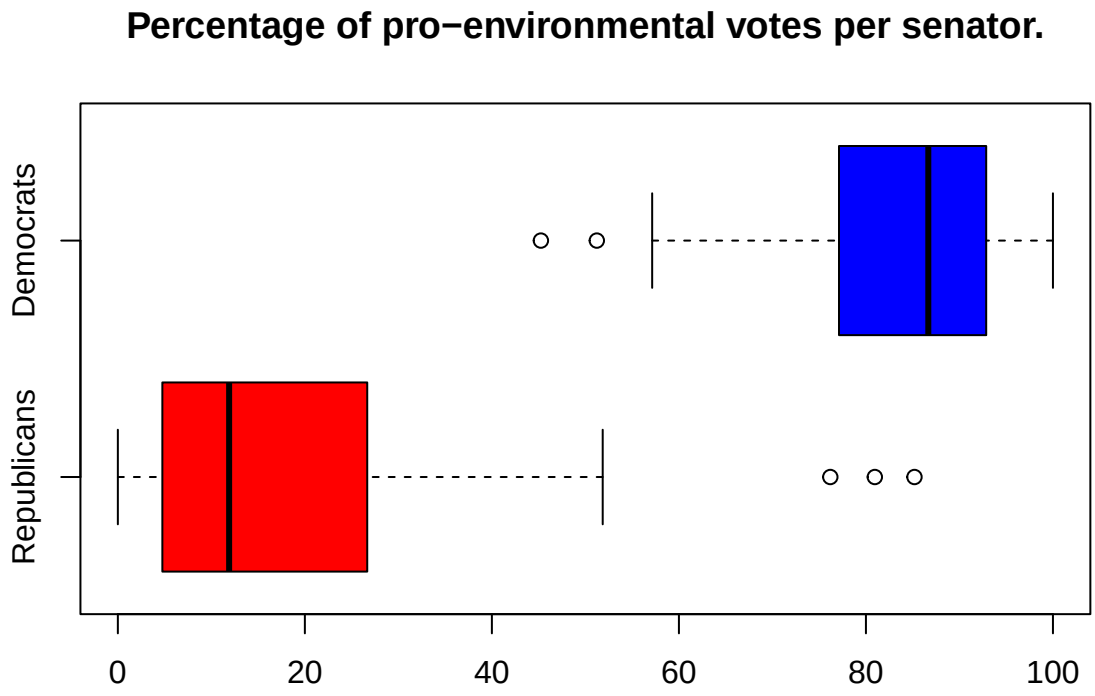
Week 02 solutions

2023-02-08

Part one: Tuesday

1. Environmental Voting of Democrats and Republicans in the U.S. Senate. In this exercise we will consider a data set according to the League of Conservation Voters of pro- and anti-environmental votes cast by U.S. senators during 2005 - 2007. Use visual representation of the data to investigate party differences in the percentage of pro-environment votes.
   Solution:

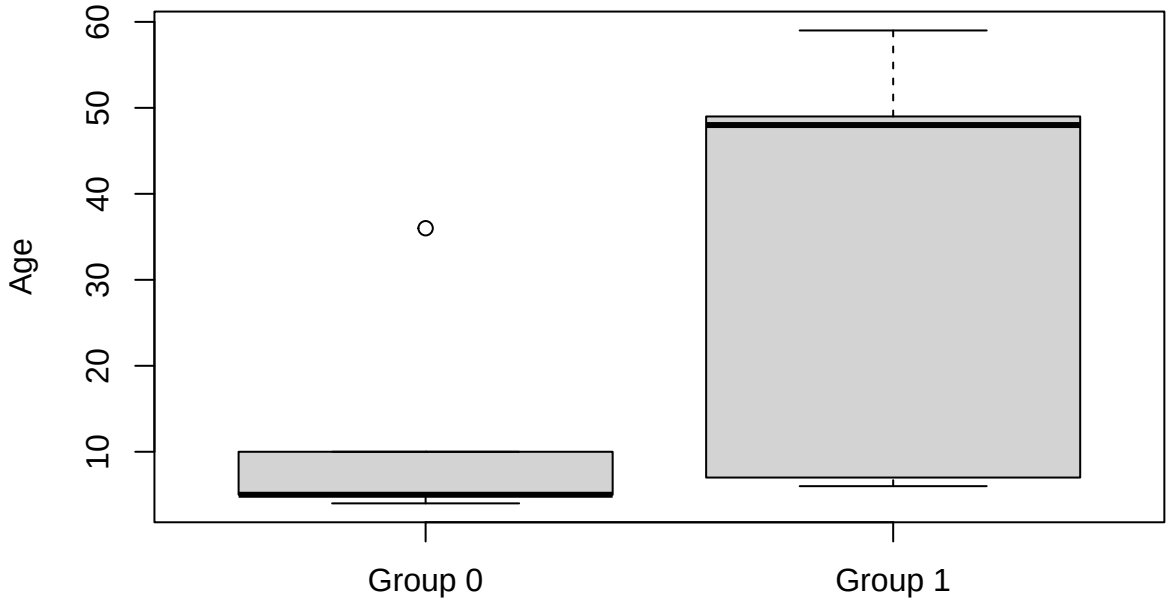
```
vote_data <- read.csv('data/Environmental votes.csv')
republicans <- dplyr::filter(vote_data, Party == "R")
democrats <- dplyr::filter(vote_data, Party == "D")
boxplot(republicans$PctPro, democrats$PctPro,
        main="Percentage of pro-environmental votes per senator.",
        names=c("Republicans", "Democrats"),
        col=c("Red", "Blue"),
        horizontal=TRUE)
```



2. Fish Oil and Blood Pressure. Consider the study from previous week about the correlation between a diet containing fish oil and blood pressure. In the study, the researchers used 7 red and 7 black playing cards to randomly the participants to the treatment groups. Does this method constitute a random sampling? Why might the results of this study be important?
   Solution:
   This depends on how the participants were randomized, whether the cards were shuffled etc. If we assume the cards were fairly shuffled, and the participants each randomly chose a card, it is a random sample.
   The results are important as they suggest consuming fish oil reduces blood pressure, which could be a potential treatment option with more research.

3. Write down the names and ages of 10 people. Using coin flips, divide them into two groups, as if for a randomized experiment. Did one group tend to get many of the older subjects? Was there anyway to predict which group would have a higher age in advance of the coin flips?
   Solution:

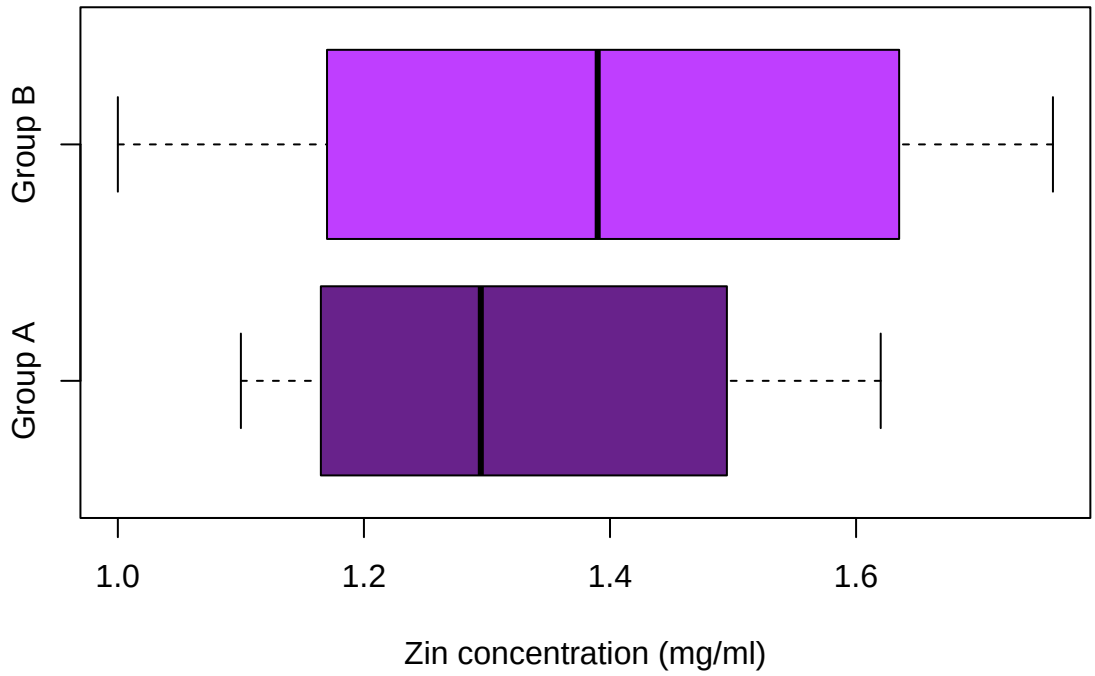
```
people <- data.frame(
  name=c("Gideon", "Maddison", "Alisha", "Brodie", "Cian", "Osman", "Ciara", "Alesha", "Jesse", "Kurtis
  age=floor(runif(10, 1, 60)),
  group=rbinom(10, 1, 0.5)
)
boxplot(
  dplyr::filter(people, group == 0)$age, dplyr::filter(people, group == 1)$age,
  names=c("Group 0", "Group 1"), ylab="Age"
)
```



It is not possible to predict which group will have a higher age, as the distribution is random.

4. Zinc concentration. An experiment was conducted where a group of rats were split into two groups. One group, group A, received a dietary supplement, while the other group, group B, did not. We are interested in the possible side effect of the treatment. In the data set, the concentration of zinc (in mg/ml) in each rat is listed, as well as the rats' treatment group. Make box-plots for each of the two groups of rats. Preferably in the same plot, and if not then at least with the scale of the axis.
Solution:

```
rat_data <- read.csv('data/Zinc concentration.csv')
boxplot(
  dplyr::filter(rat_data, Group == "A")$Zinc, dplyr::filter(rat_data, Group == "B")$Zinc,
  names=paste("Group", c("A", "B")), xlab="Zin concentration (mg/ml)",
  horizontal=TRUE,
  col=c("darkorchid4", "darkorchid1")
)
```



Part two: Thursday

1. Explain the difference between regression, regression model, and simple linear regression model.
Solution:

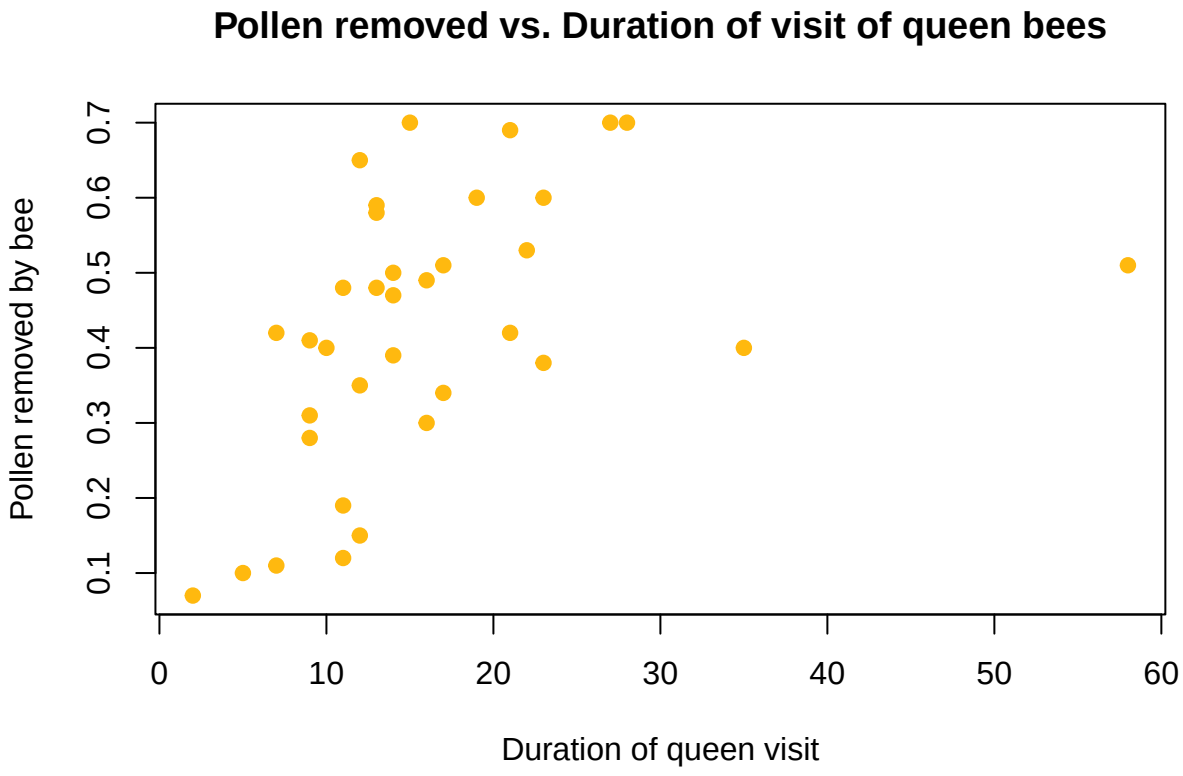
- Regression: Going back to a previous state or a technique for prediction of a relation between two variables.

- Regression model: The mathematical model used in the regression.
- Simple linear regression model: The mathematical model used by simple linear regression. Has the form $\hat{Y}_i = \alpha + \beta X_i$ where $\hat{Y}_i$ is the Predicted value, X_i is the Predictor value, α is the intercept and β is the slope.

2. Pollen Removal. As part of a study to investigate reproductive strategies in plants, biologists recorded the time spent at sources of pollen and the proportions of pollen removed by bumblebee queens and honeybee workers pollinating a species of lily (Data from L. D. Harder and J. D. Thompson, "Evolutionary Options for Maximizing Pollen Dispersal of Animal-pollinated Plants", American Naturalist 133 (1989): 323-44.). Find the data in this weeks folder "data sets".

(a) Draw a scatterplot of proportion of pollen removed versus duration of visit, for the bumblebee queens. *Solution*:

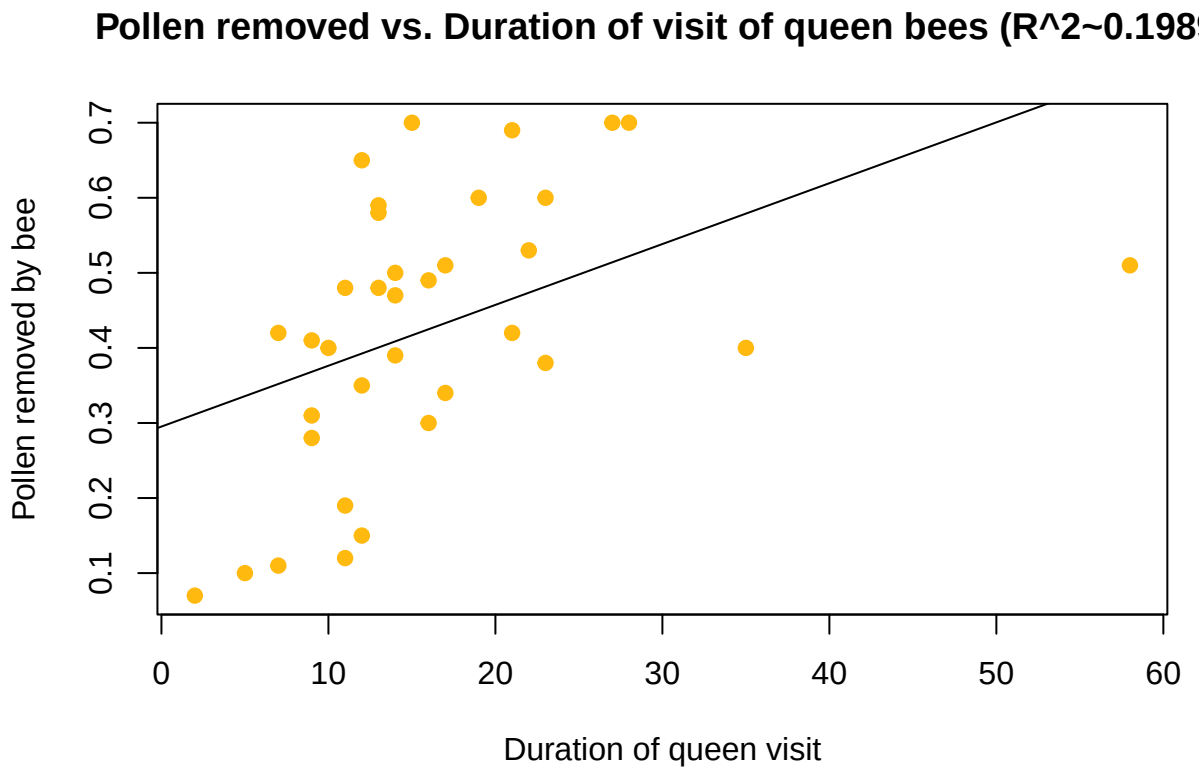
```
bee_data <- read.csv('data/Pollen Removal.csv')
queens_data <- dplyr::filter(bee_data, BeeType == "Queen")
plot(queens_data$DurationOfVisit, queens_data$PollenRemoved,
     main="Pollen removed vs. Duration of visit of queen bees",
     xlab="Duration of queen visit", ylab="Pollen removed by bee",
     col="darkgoldenrod1", pch=19)
```



(b) Find and follow a guide on how to make a simple linear regression line with R.

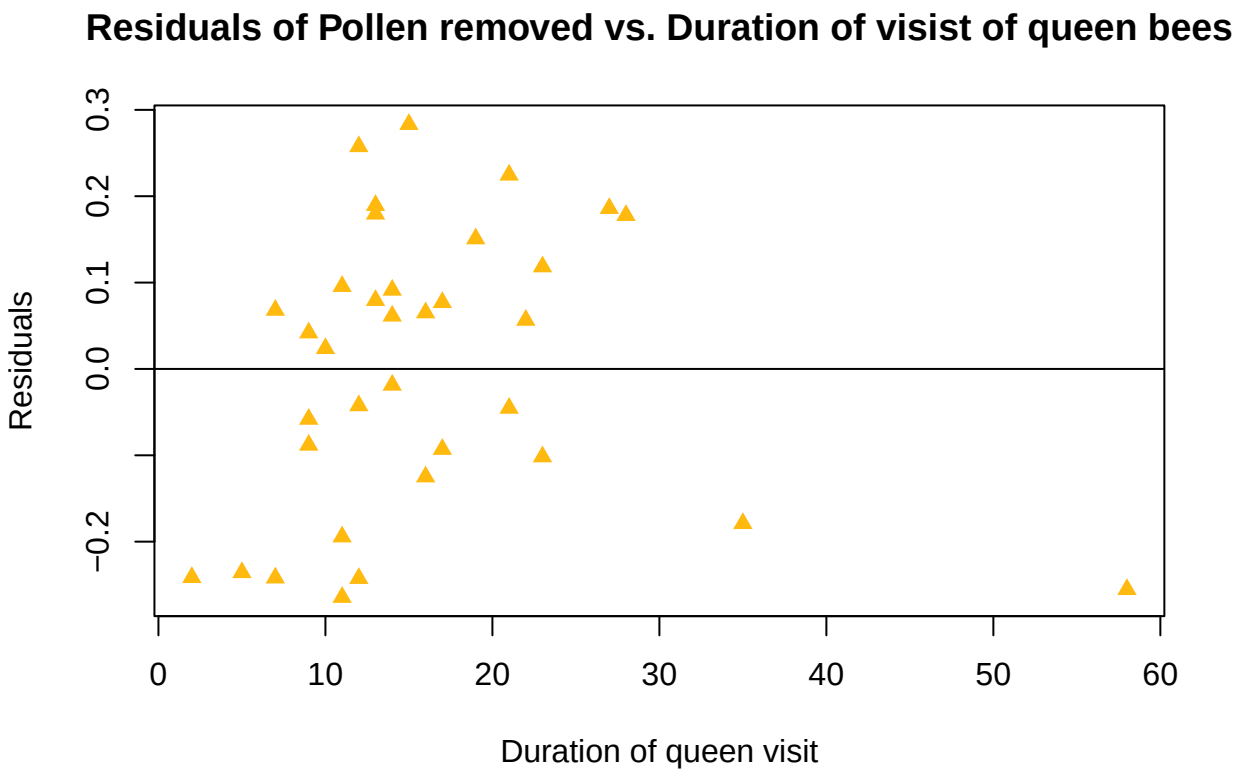
(c) Draw the estimated regression line on the scatterplot.
Solution:

```
queens_relation <- lm(PollenRemoved ~ DurationOfVisit, data=queens_data)
R_squared <- summary(queens_relation)$r.squared
plot(queens_data$DurationOfVisit, queens_data$PollenRemoved,
     main=sprintf("Pollen removed vs. Duration of visit of queen bees (R^2~%s)", round(R_squared, digits
     xlab="Duration of queen visit", ylab="Pollen removed by bee",
     col="darkgoldenrod1", pch=19)
abline(queens_relation)
```



(d) What problems are evident in the residual plot?
Solution:
First we have to draw the residual plot.

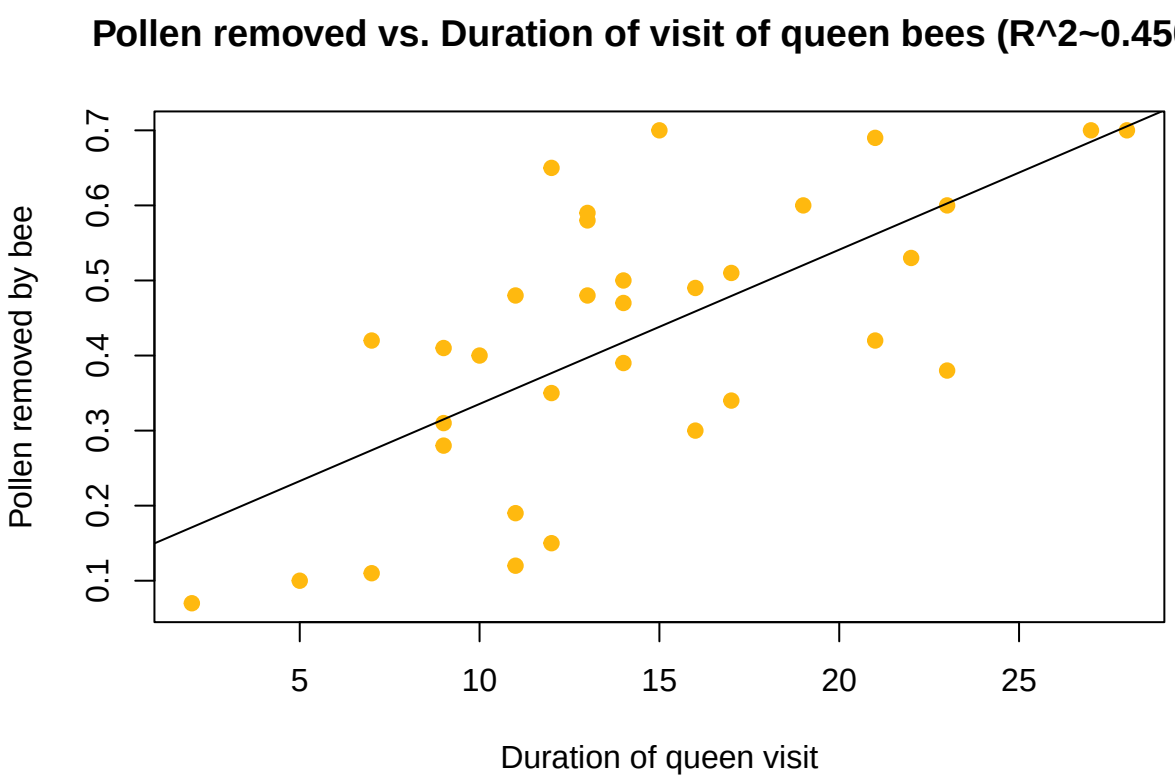
```
queens_residual <- resid(queens_relation)
plot(queens_data$DurationOfVisit, queens_residual,
     ylab="Residuals", xlab="Duration of queen visit",
     main="Residuals of Pollen removed vs. Duration of visist of queen bees",
     col="darkgoldenrod1", pch=17)
abline(0, 0)
```



It is clear that the two extreme values > 30 have a very high influence of the linear regression line. Without them the cluser on the left may have a line with a better fit.

(e) Try fitting the regression only for those times less than 31 seconds (i.e. excluding the two longest times). Does this fit better?
Solution:

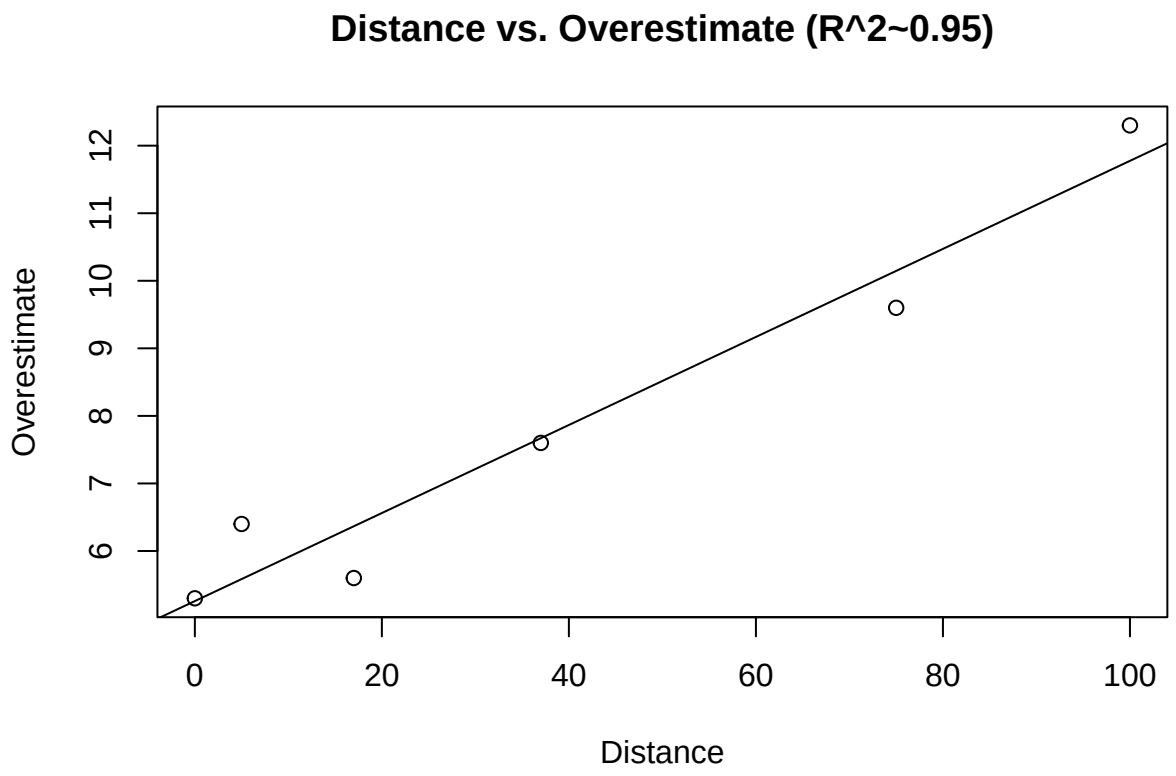
```
queens_filt <- dplyr::filter(queens_data, DurationOfVisit < 31)
queens_relation_upd <- lm(PollenRemoved ~ DurationOfVisit, data=queens_filt)
R_squared <- summary(queens_relation_upd)$r.squared
plot(queens_filt$DurationOfVisit, queens_filt$PollenRemoved,
     main=sprintf("Pollen removed vs. Duration of visit of queen bees (R^2~%s)", round(R_squared, digits
     xlab="Duration of queen visit", ylab="Pollen removed by bee",
     col="darkgoldenrod1", pch=19)
abline(queens_relation_upd)
```



As we can see, the R^2 value has almost doubled, which means this linear regression is a much better predictor.

3. Sampling Bias in Exit Polls. Exit pollsters predict election results before final counts are tallied by sampling voters leaving voting locations. The pollsters have no way of selecting a random sample, so they instruct their interviewers to select every third exiting voter, or fourth, or fifth, or tenth, or some other specific number. Some voters refuse to participate or avoid the interviewer. If the refusers and avoiders have the same voting patterns as the rest of the population, then it shouldn't matter; the sample, although not random, wouldn't be biased. If, however, one candidate's voters are more likely to refuse or avoid interview, the sample would be biased and could lead to misleading conclusion.
   On November 4, 2004, exit pollsters incorrectly predicted that John Kerry would win the U.S. presidential election over George W. Bush. The exit polls overstated the Kerry advantage in 42 of 50 states. No one expects exit polls to be exact, but chances alone cannot reasonably explain this discrepancy. Although fraud is a possibility, the data are also consistent, with Bush supporters being more likely than Kerry supporters to refuse or avoid participation in the exit poll.
   In a postelection evaluation, the exit polling agency investigated voter avoidance of interviewers. The data set includes the average Kerry exit poll overestimate (determined after the actual counts were available for a large number of voting precincts, grouped according to the distance of the interviewer from the door. If Bush voters were more likely to avoid interviewers in general, one might also expect a greater avoidance with increasing distance to the interviewer (since there is more opportunity to escape). A positive relationship between distance of the interviewer from the door and amount of Karry overestimate, therefore, would lend credibility to the theory that Bush voters were more likely to avoid exit poll interviews. How strong is the evidence that the mean Karry overestimate increases with increasing distance of interviewer from the door? (Data from Evaluation of Edison/ Mitofsky Election System 2004 prepared by Edison Media Research and Mitofsky International for the National Election Pool (NEP), January 15, 2005)
   Solution:

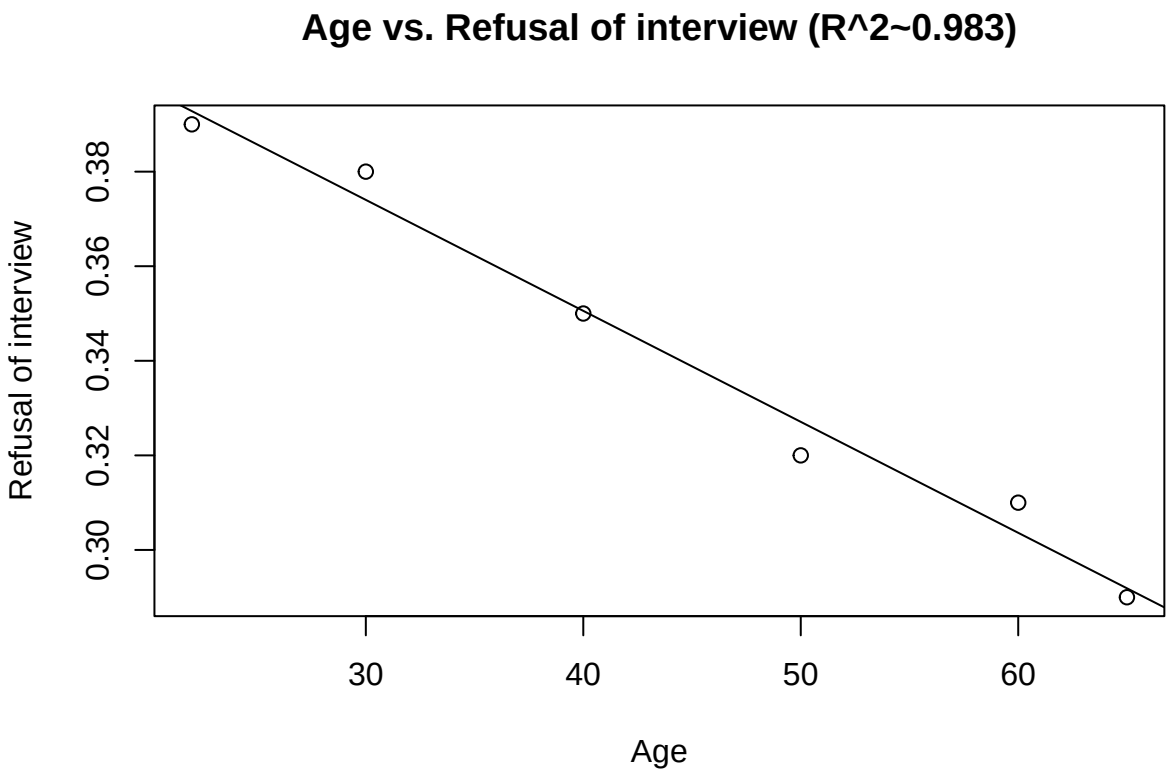
```
sampling_bias <- read.csv('data/Sampling Bias.csv')
sampling_relation <- lm(OverEstimate ~ Distance, data=sampling_bias)
R_squared <- summary(sampling_relation)$r.squared
plot(sampling_bias$Distance, sampling_bias$OverEstimate,
     main=sprintf("Distance vs. Overestimate (R^2~%s)", round(R_squared, 3)),
     xlab="Distance", ylab="Overestimate")
abline(sampling_relation)
```



The R^2 value is almost 1, which means the regression is a very good predictor of values. It is however hard to draw conclusions on this data, as there are very few datapoints.

4. Sampling Bias in Exit Polls II. This exercise is about differential interview refusal rates in the exit polls conducted in the 2004 U.S. presidential election. The data shows the average proportion of voters who refuse to be interviewed at precincts grouped according to the approximate age of the interviewer. What evidence do the data provide that the mean refusal rate decreased with increasing age of interviewer?
   Solution:

```
sampling_bias <- read.csv('data/Bias II.csv')
sampling_relation <- lm(Refusal ~ Age, data=sampling_bias)
R_squared <- summary(sampling_relation)$r.squared
plot(sampling_bias$Age, sampling_bias$Refusal,
     main=sprintf("Age vs. Refusal of interview (R^2~%s)", round(R_squared, 3)),
     xlab="Age", ylab="Refusal of interview")
abline(sampling_relation)
```



Again, R^2 is very high which shows the line as an almost perfect fit, but the amount if data points are very limited so it is hard to draw conclusions.