Michael Spainhour

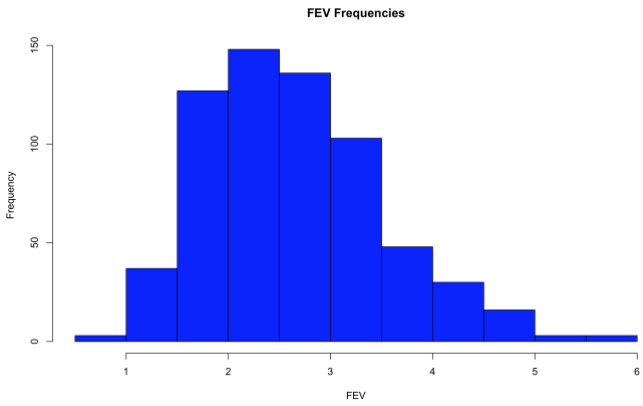
Scientific Computing

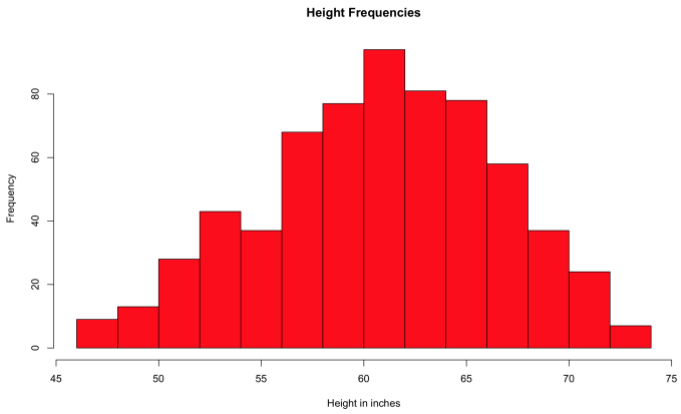
Lab 1

Before diving too much into the data, I decided to look at the frequencies of values for ‘age’ and ‘fev’. I noticed that the data is pretty normally distributed for both variables. This does not tell us much from the start, but we can see the way the frequencies of fev and height behave. For the FEV frequencies, from 0-2 FEV, there is a steep increase in frequency. I predict that this is the case because some of the lower FEVs are outliers in the data and perhaps they represent an indicator for asthma or another pulmonary disease.

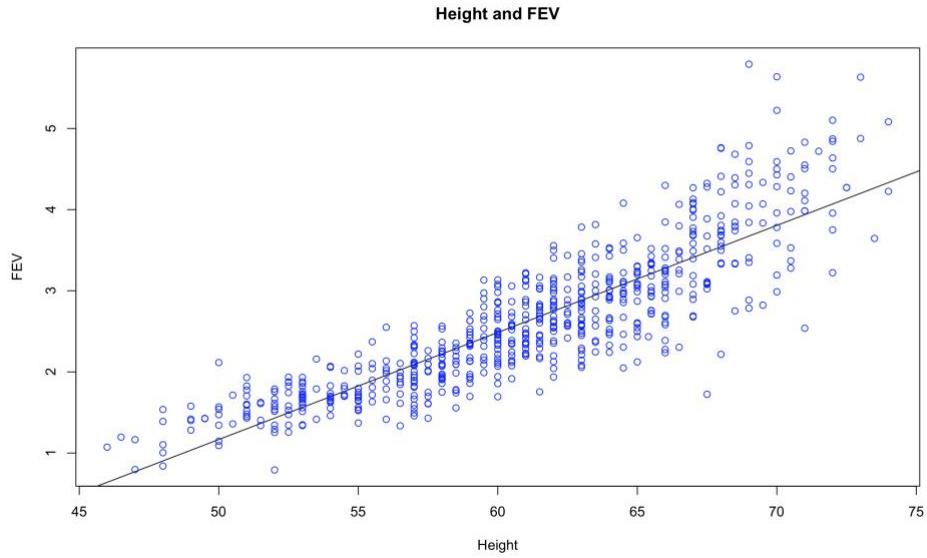
hist(dataset$fev, main='FEV Frequencies', xlab='FEV', col='blue')

hist(dataset$height, main='Height Frequencies', xlab='Height in inches', col='red')





Is there a correlation between height and FEV? I used the correlation function *cor(dataset$height, dataset$fev)* and got .868. With this number being as close as it is to one, we can say there is a fairly strong correlation between height and FEV. Which makes sense because as one gets taller, they should be able to expel more and more air.



Here is a look at the scatter plot I created with a linear regression line.

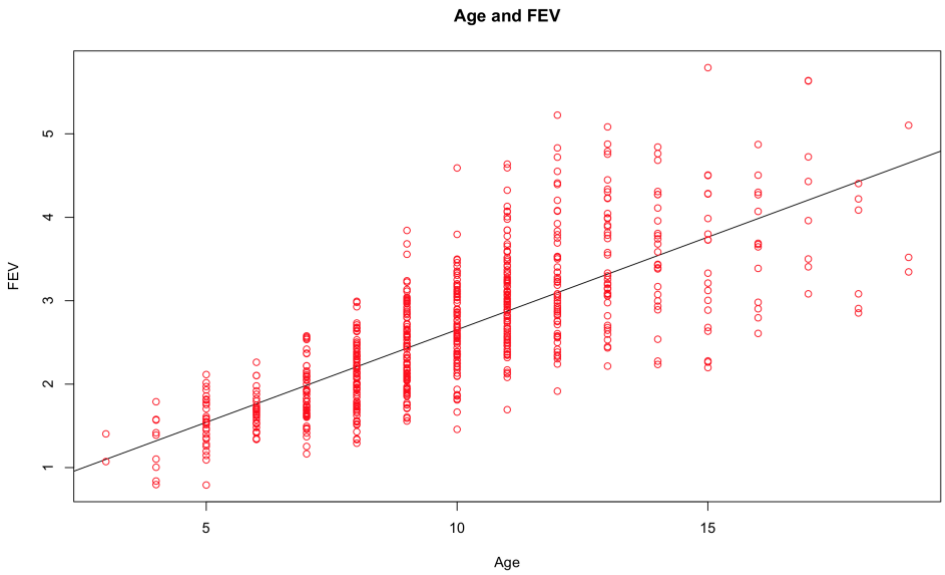
plot(dataset$height, dataset$fev, main='Height and FEV', abline(lm(fev~height, data=dataset)), xlab='Height', ylab='FEV')

We can see points in the data where there are values pretty far below the line. These values are much below their predicted values. There are even some values far above the linear regression line. This supports our finding for the correlation value.

I also took a look at the relationship between age and FEV.

cor(dataset$age, dataset$fev)

Going through the same process as before, I got a correlation value of .756. This is also to be expected because as one gets older and more mature, they should be able to expel more air.



This scatter plot, similar to our scatter plot for height and FEV, shows a fair correlation between age and FEV. Although it is not quite as strong, we can see that the general trend shows a relationship between increasing age and increasing FEV.

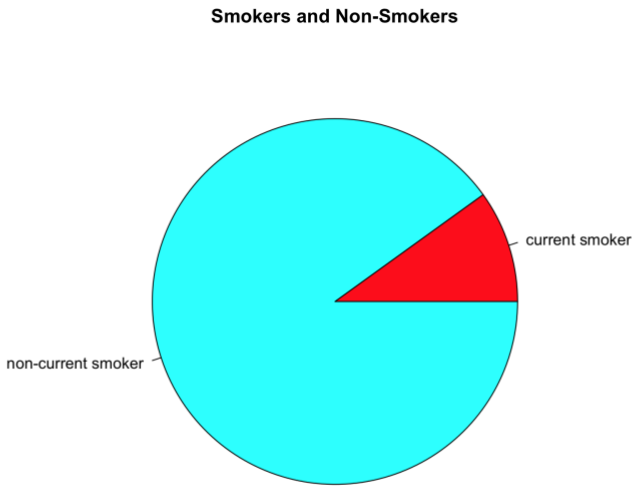
plot(dataset$age, dataset$fev, main='Age and FEV', abline(lm(fev~age, data=dataset)), xlab='Age', ylab='FEV', col='red')

It does not seem like there are too many cases in the data where an FEV level might be too low to be noteworthy for a specific person, and I wondered if being a smoker had an impact on some of those numbers. Here I got the frequency count of smokers and non-smokers in the dataset, and created a pie chart from it for visualization.

smoker_nonsmoker_freqs <- table(dataset$smoke)

smoker_dataframe <- data.frame(smokers=c('current smoker', 'non-current smoker'), freqs=c(smoker_nonsmoker_freqs[["current smoker"]][1], smoker_nonsmoker_freqs[["non-current smoker"]][1]))

pie(smoker_dataframe$freqs, smoker_dataframe$smokers, main='Smokers and Non-Smokers', col=rainbow(2))



For our dataset, I found that there were way more non-current smokers than there were current smokers so it might be hard to tell if smoking actually has a correlation with FEV values. I found the correlation between ‘smoke’ and ‘fev’ to be .245. This is lower than I would have estimated it to be. If there had been more current-smokers in the dataset, perhaps we would have seen something different.

```
fev_and_smoke <- data.frame(dataset$fev, dataset$smoke, stringsAsFactors=FALSE)
fev_and_smoke$dataset.smoke[fev_and_smoke$dataset.smoke=="current smoker"] <- 1
fev_and_smoke$dataset.smoke[fev_and_smoke$dataset.smoke=="non-current smoker"]
<- 0
 fev_and_smoke$dataset.smoke <-
as.numeric(as.character(fev_and_smoke$dataset.smoke))
cor(fev_and_smoke$dataset.fev, fev_and_smoke$dataset.smoke)
```

I created a new data frame with two columns: one with FEV levels, and one with the corresponding smoker status. I changed the values for current smokers to 1, and the values for non-current smokers to 0.