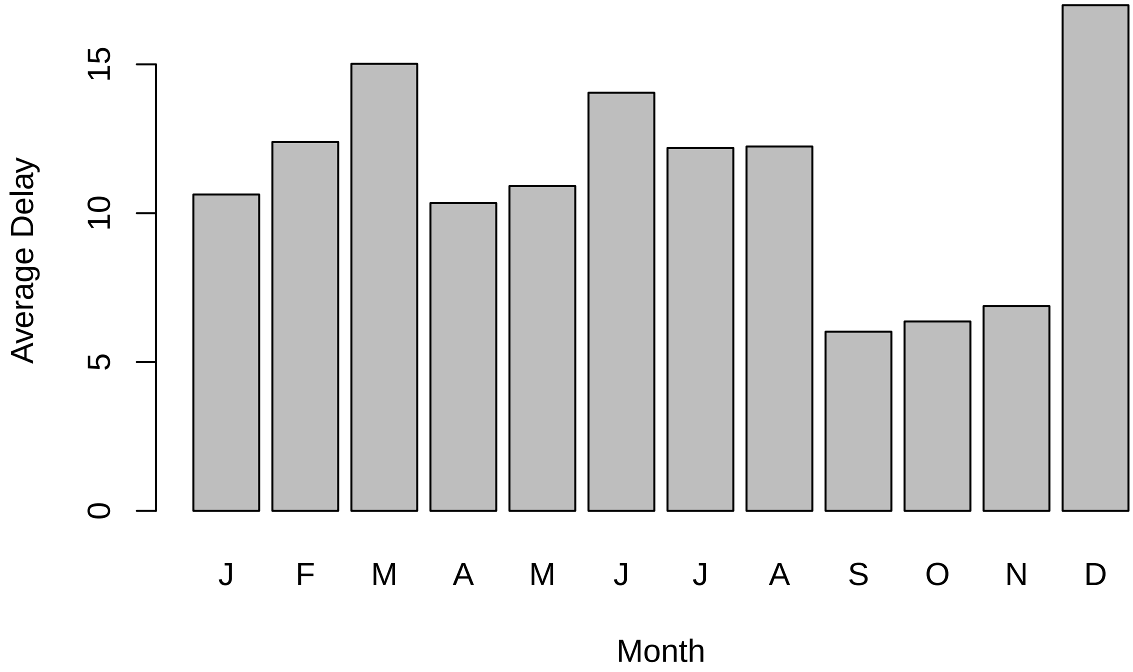
Exercise 1

ABIA flights

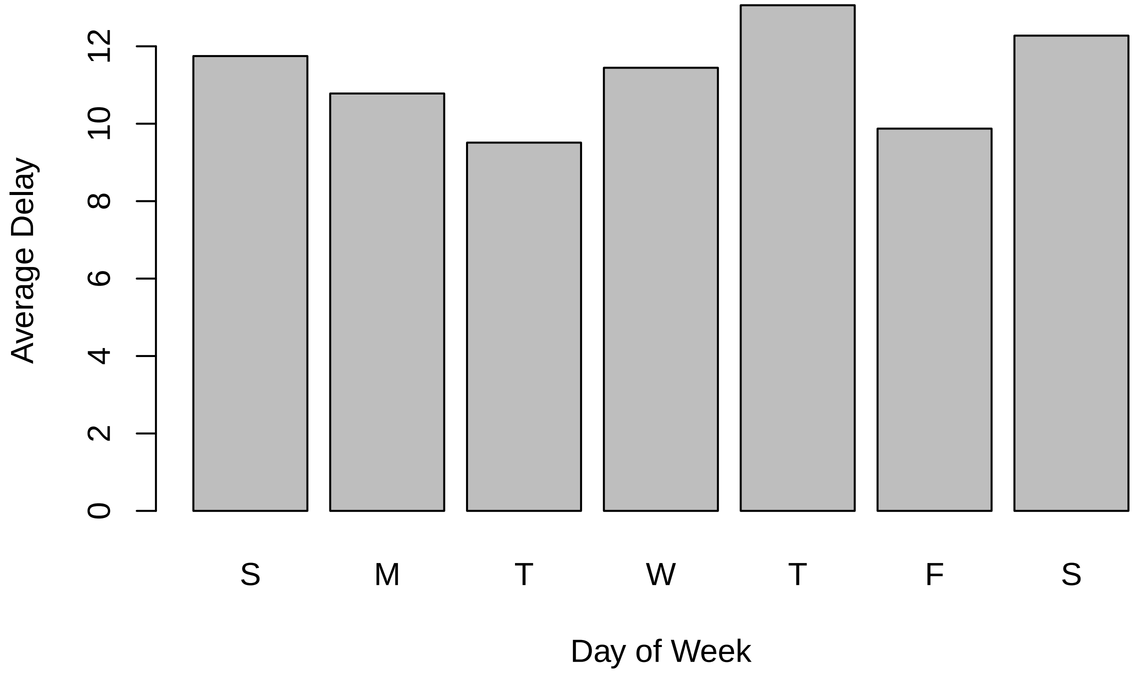
We analyze the average delay in departures by month and day of week, where the delay is counted as 0 if the plane departs early. We ignore records in which the delay is NA.

```
flights <- read.csv("~/Documents/SDS323Assignments/ABIA.csv")
months = flights$Month
delays = flights$DepDelay
totalDelays = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
totals = c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
avgs = c()
for (i in 1:length(months)){
  if(!is.na(delays[i])){
  totalDelays[months[i]] = totalDelays[months[i]] + max(delays[i], 0)
  totals[months[i]] = totals[months[i]] + 1
  }
}
for (i in 1:12){
  avgs[i] = totalDelays[i] / totals[i]
}
months = c('J', 'F', 'M', 'A', 'M', 'J', 'J', 'A', 'S', 'O', 'N', 'D')
barplot(avgs, names.arg = months, xlab = 'Month', ylab = 'Average Delay')
```



We observe that on average there tends to be a spike in mean delays in December, but smaller delays in September through November.

```
dayOfWeek = flights$DayOfWeek
delays = flights$DepDelay
totalDelays = c(0, 0, 0, 0, 0, 0, 0)
totals = c(0, 0, 0, 0, 0, 0, 0)
avgs = c()
for (i in 1:length(dayOfWeek)){
  if(!is.na(delays[i])){
  totalDelays[dayOfWeek[i]] = totalDelays[dayOfWeek[i]] + max(delays[i], 0)
  totals[dayOfWeek[i]] = totals[dayOfWeek[i]] + 1
  }
}
for (i in 1:7){
  avgs[i] = totalDelays[i] / totals[i]
}
dayOfWeek = c('S', 'M', 'T', 'W', 'T', 'F', 'S')
barplot(avgs, names.arg = dayOfWeek, xlab = 'Day of Week', ylab = 'Average Delay')
```



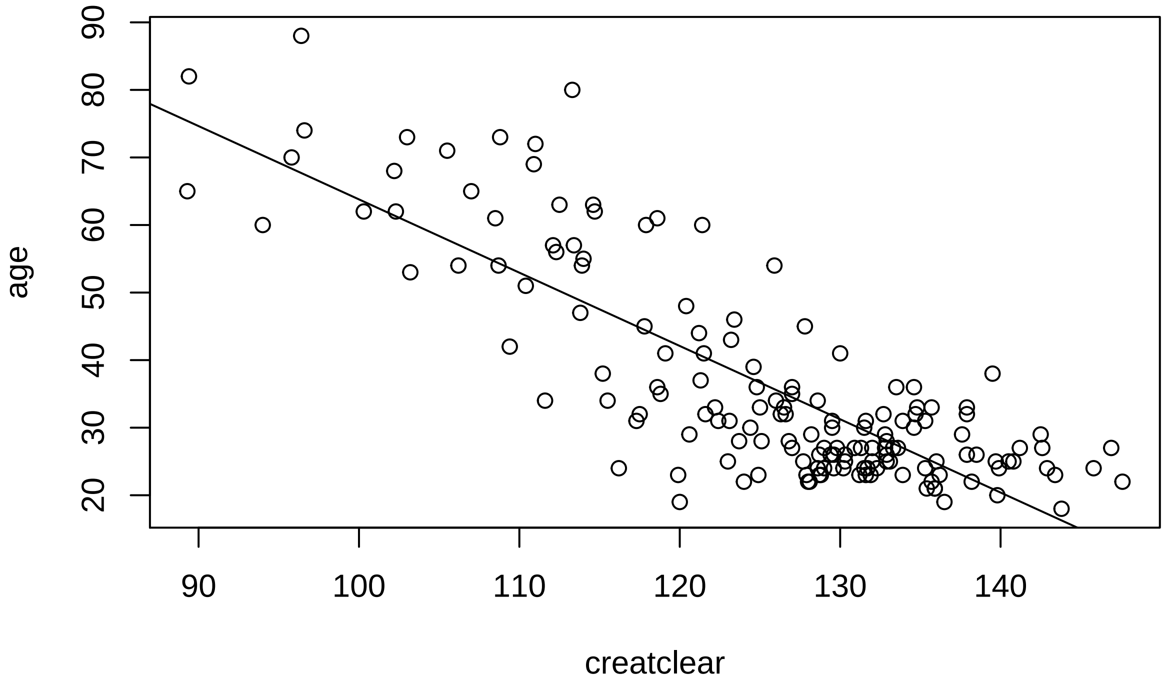
There does not seem to be a major correlation between day of the week and delays

Regression Practice

```
creatinine <- read.csv("~/Documents/SDS323Assignments/creatinine.csv")
lm.fit = lm(age~creatclear, data = creatinine)
intercept = summary(lm.fit)$coefficients[1,1]
slope = summary(lm.fit)$coefficients[2,1]
```

Plot:

```
plot(age~creatclear, data = creatinine)
abline(intercept, slope)
```



Estimated creatinine clearance rate for 55-year-old:

```
intercept + slope * 55
```

```
## [1] 112.6141
```

Creatinine change with age:

```
slope
```

```
## [1] -1.084897
```

On average, we expect that, for each additional year of age, creatinine clearance rate decreases by approximately 1.085 mL/min.

Expected clearance rate for a 40-year-old:

```
act40 = 135
exp40 = intercept + slope * 40
```

Difference:

```
act40 - exp40
```

```
## [1] 6.112475
```

Expected clearance rate for a 60-year-old:

```
act60 = 112
exp60 = intercept + slope * 60
```

Difference:

```
act60 - exp60
```

```
## [1] 4.810408
```

We conclude that, based on our linear model, the 40-year-old appears to be healthier for his age, because of his higher adjusted creatinine.

Green Buildings

```
data <- read.csv("~/Documents/SDS323Assignments/greenbuildings.csv")
dataFiltered = filter(data, leasing_rate >= 10)
green = filter(dataFiltered, green_rating == 1)
nonGreen = filter(dataFiltered, green_rating == 0)
```

We attempt to look for confounding variables:

```
mean(nonGreen$age)
```

```
## [1] 49.30808
```

```
mean(green$age)
```

```
## [1] 23.88012
```

We observe that green houses are, on average, much younger than non-green houses. We add this variable to our regression.

```
lm.fit = lm(Rent~age+green_rating, data = data)
```

We already observe that, with the age variable, the green rating effect drops drastically, suggesting that age is a major confounding variable. It also appears that older houses tend to charge less rent. Suppose we consider houses designed in the past 30 years, which will apply to the newly designed house.

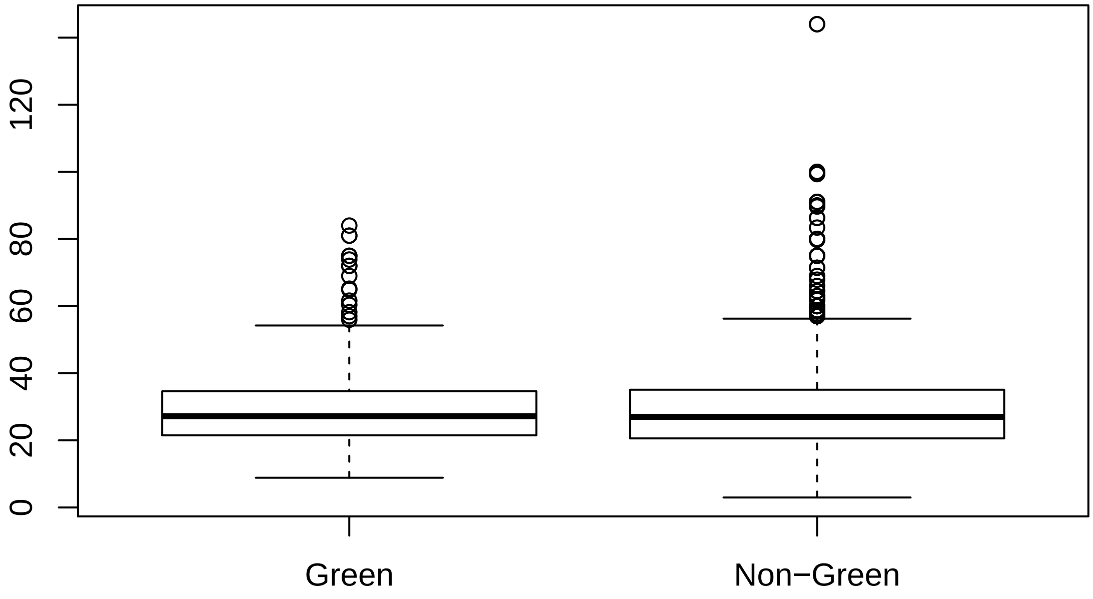
```
recentNonGreen = filter(nonGreen, age <= 30)
recentGreen = filter(green, age <= 30)
summary(recentNonGreen$Rent)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.98   20.58   27.00   29.23   35.10  144.00
```

```
summary(recentGreen$Rent)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.87   21.49   27.20   29.29   34.63   84.00
```

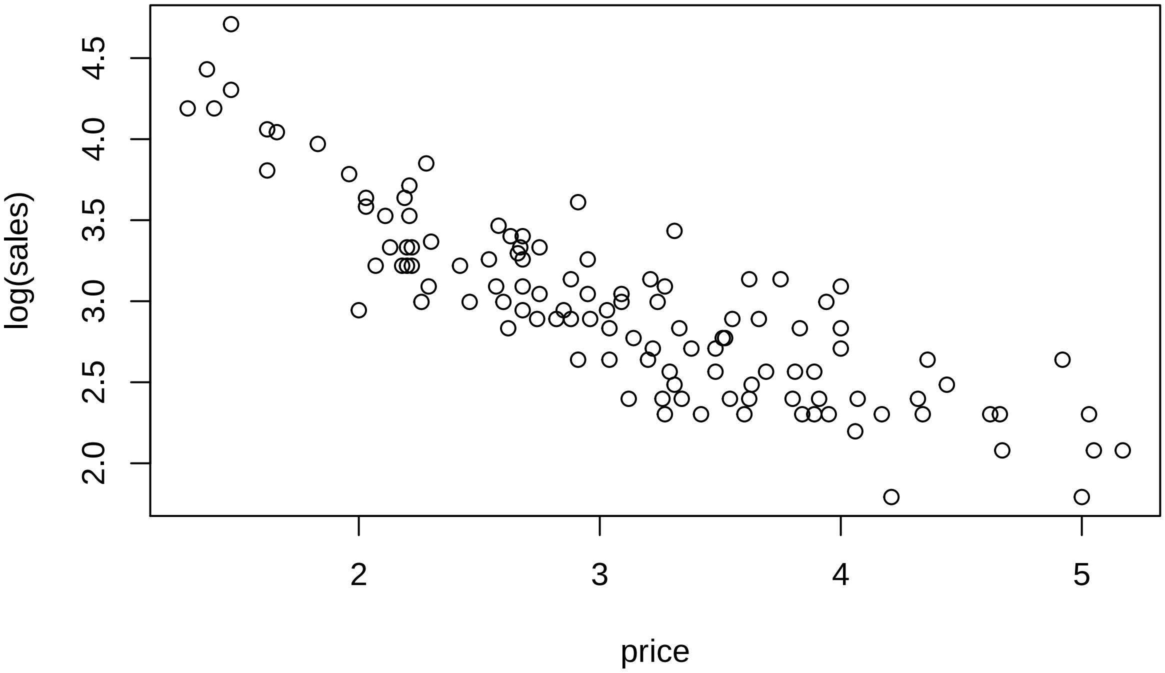
```
boxplot(recentGreen$Rent, recentNonGreen$Rent, names = c('Green', 'Non-Green'))
```



The medians are now virtually indistinguishable, with a difference of 0.2. For the next 30 years, the profit for going green, according to this model, would be approximately 250,000 x 0.2 x 30 = 1.5 million, well below our initial 5 million dollar investment. To summarize, the staff member failed to account for the age of the house as a confounding factor when performing his analysis, and the fact that the house will be brand new. When we only analyze the houses built in the last 30 years, we see that the additional profit for choosing the green option essentially vanishes, calling into question the assertion that the green option is profitable after less than 10 years.

Milk Prices

```
data <- read.csv("~/Documents/SDS323Assignments/milk.csv")
plot(log(sales)~price, data = data)
```



We observe that the data is non-linear and therefore fit the logarithms.

```
lm.fit = lm(log(sales)~log(price), data = data)
lm.fit
```

```
##
## Call:
## lm(formula = log(sales) ~ log(price), data = data)
##
## Coefficients:
## (Intercept)   log(price)
##       4.721       -1.619
```

We get the following equation, where x is the price per carton and y is the expected number of cartons sold:

$y = \alpha x^\beta$ where $\alpha = e^{4.721} = 71.6$ and $\beta = -1.619$

Our overall expected profit is then the product of the profit per carton, $x - 1$, and the expected number of cartons sold:

$f(x) = \alpha x^\beta (x-1) = \alpha(x^{\beta+1} - x^\beta)$

Taking the derivative and setting to zero:

$0 = f'(x) = \alpha((\beta+1)x^{\beta} - \beta x^{\beta-1})$

Solving, we get the optimal price: $x = \frac{\beta}{\beta+1} = \frac{-1.619}{-0.619} = \$2.62$