# HW #1

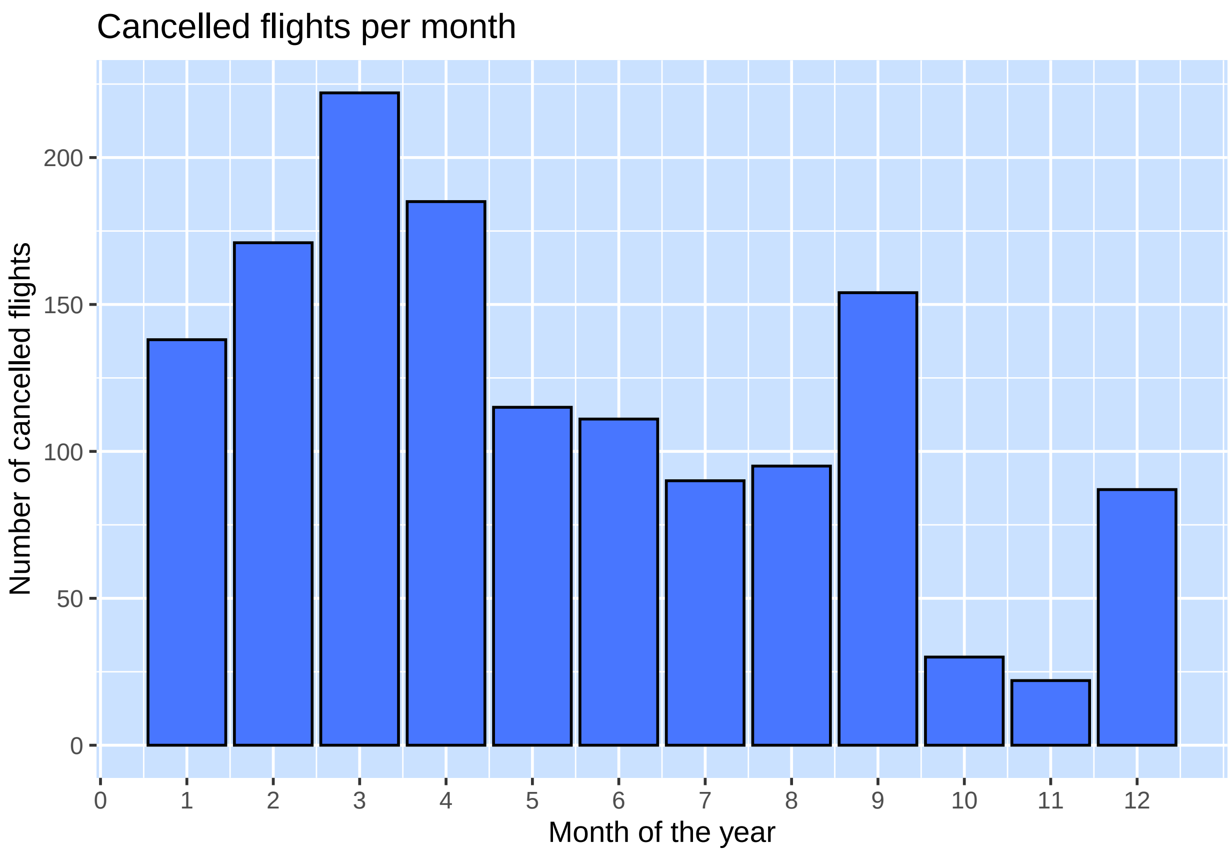
Rylan, Sam

2/11/2020

### 1. Data visualization: flights at ABIA

## Which month has the most cancelled flights?

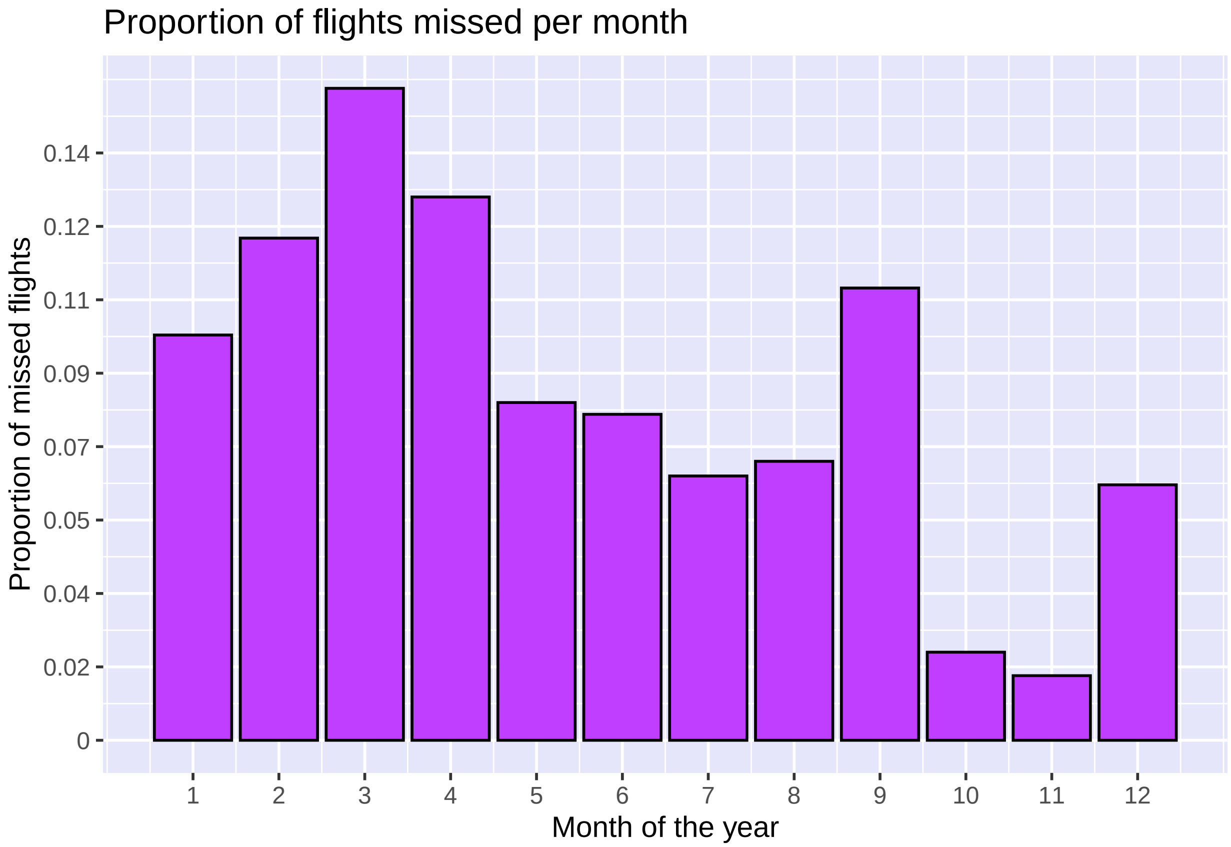
```
ABIA <- read.csv("C:/Users/shigg/OneDrive/Documents/College Notes/R stuff/ABIA.txt")

library(ggplot2)

cflight <- (ABIA[ABIA$Cancelled == 1 , ]$Month)
ggplot(ABIA[ABIA$Cancelled == 1 , ], aes(x = Month))+
  geom_bar(color = "black", fill = "royalblue1") +
  scale_x_continuous(breaks = 0 : 12) +

  ggtitle("Cancelled flights per month") +
  xlab("Month of the year") +
  ylab("Number of cancelled flights") +
  theme(panel.background = element_rect(fill = "lightsteelblue1", colour = "lightsteelblue1", size = 0.5
```



## From the bar graph, one can see that March, the third month of the year, has the most number of missed flights.

### During which month is it the worse travel when considering proportion of flights that are cancelled per month?

```
bflight <- (ABIA$Cancelled == 1)
ggplot(ABIA[ABIA$Cancelled == 1 , ], aes(x = Month))+
  ggtitle("Proportion of flights missed per month") +
  xlab("Month of the year") +
  ylab("Proportion of missed flights") +
  geom_bar(color = "black", fill = "darkorchid1") +
  scale_y_continuous(breaks = seq(0, 200 , 25), labels = round( seq(0, 200 , 25 )/ sum(bflight), 2))+
  scale_x_continuous(breaks = 1 : 12) +
  theme(panel.background = element_rect(fill = "lavender", colour = "lavender", size = 0.5))
```

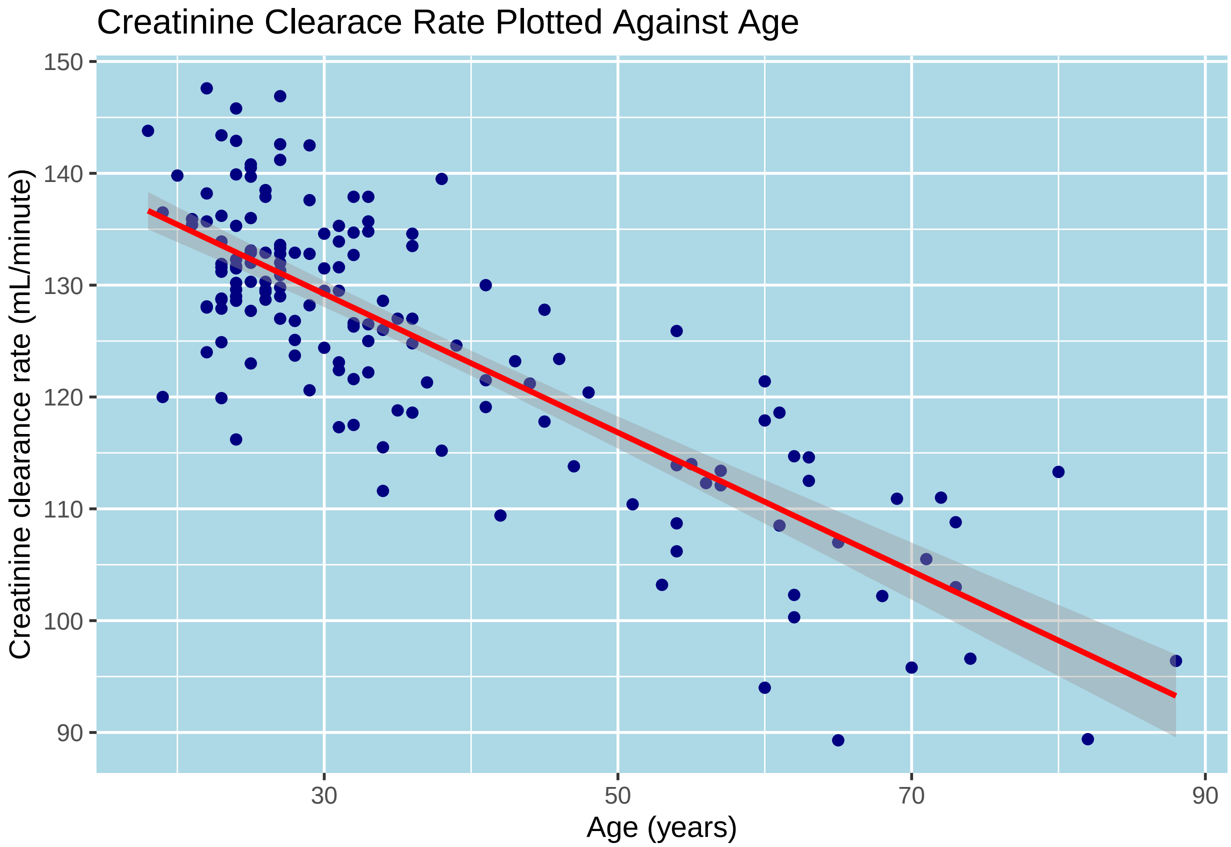


**From this bargraph one can tell that, even when compared proportionally, March still has the the most cancelled flights. Thus, we conclude that the worse month to travel during is March.**

### 2. Regression practice

### 2.1 What creatinine clearance rate should we expect, on average, for a 55-year-old?

```
creatinine <- read.csv("C:/Users/shigg/OneDrive/Documents/College Notes/R stuff/creatinine.txt")

ggplot(data = creatinine, aes(x = age, y = creatclear)) +
  theme(legend.position = "none") +
  geom_point(color = "navy") +
  ggtitle("Creatinine Clearace Rate Plotted Against Age")+
  xlab("Age (years)") +
  ylab("Creatinine clearance rate (mL/minute)") +
  geom_smooth(method = "lm", color = "red") +
  theme(panel.background = element_rect(fill = "lightblue", colour = "lightblue", size = 0.5))
```



## Just from looking at the scatterplot and regression line, we can guess that a 55 year old to have a creatinine clearance rate a little bit below 115 mL/minute.

```
lm1 <- lm(data = creatinine, creatclear ~ age)
new_data = data.frame(age = 55)
predict(lm1, new_data)
```

```
##       1
## 113.723
```

## Using the regression line, we found that the predicted creatinine clearance rate for a 55 year old is 113.723 mL/minute.

### 2.2 How does creatinine clearance rate change with age? (This should be a number with units ml/minute per year.)

```
coef(lm1)
```

```
## (Intercept)         age
## 147.8129158  -0.6198159
```

# The slope of our regression line, -0.6198, tells us that for every increase in one year of age, creatinine clearance decreases by about 0.62 mL/minute.

### 2.3 Whose creatinine clearance rate is healthier (higher) for their age: a 40-year-old with a rate of 135, or a 60-year-old with a rate of 112

```r
x1 <- 40
y1 <- 135
x2 <- 60
y2 <- 112

percent_change <- function(given_x_value, given_y_value) {
  new_data1 = data.frame(age = given_x_value)
  predicted_y_value <- predict(lm1, new_data1)
  #predicted_y_value <- (147.813 - (0.6198 * given_x_value))
  percent_change_value <- ((given_y_value - predicted_y_value) / predicted_y_value)
  final_percent <- percent_change_value * 100
  return(final_percent)
}

percent_change(x1, y1)
```

```
##        1
## 9.738003
```

```r
percent_change(x2, y2)
```

```
##        1
## 1.243885
```

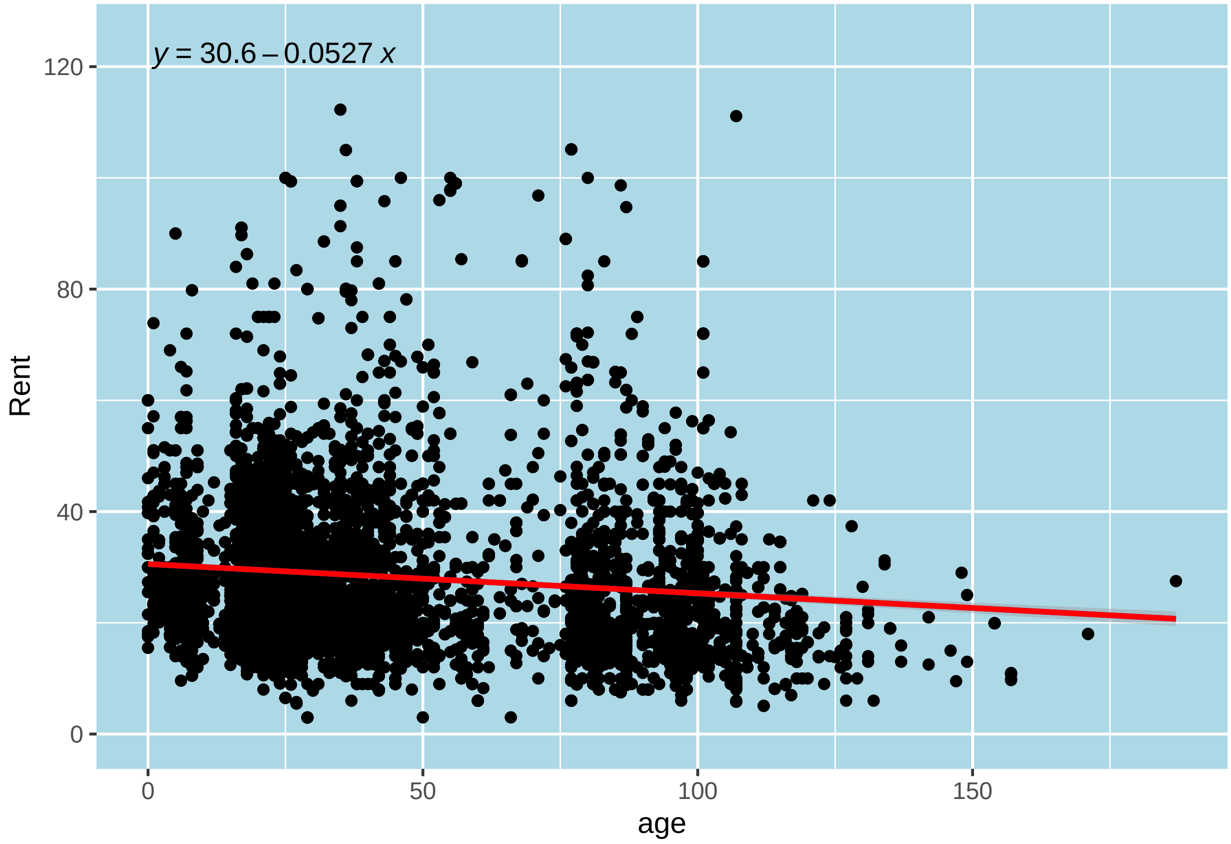
# Both the 40 year old with a rate of 135 and the 60 year old with a rate of 112 have a higher than predicted creatinine clearance rate, so they are both healthier than expected for their age. However, the 40 year old has an actual rate 9.74% above the predicted rate, whereas the 60 year old only has an actual rate 1.24% above the predicted rate, so the 40 year old is healthier for his age.

#### 3. Green buildings

# He neglected to take confounding variables, such as age, into consideration.

# Is age a counfounding variable that correlates with rent and might affect profit?

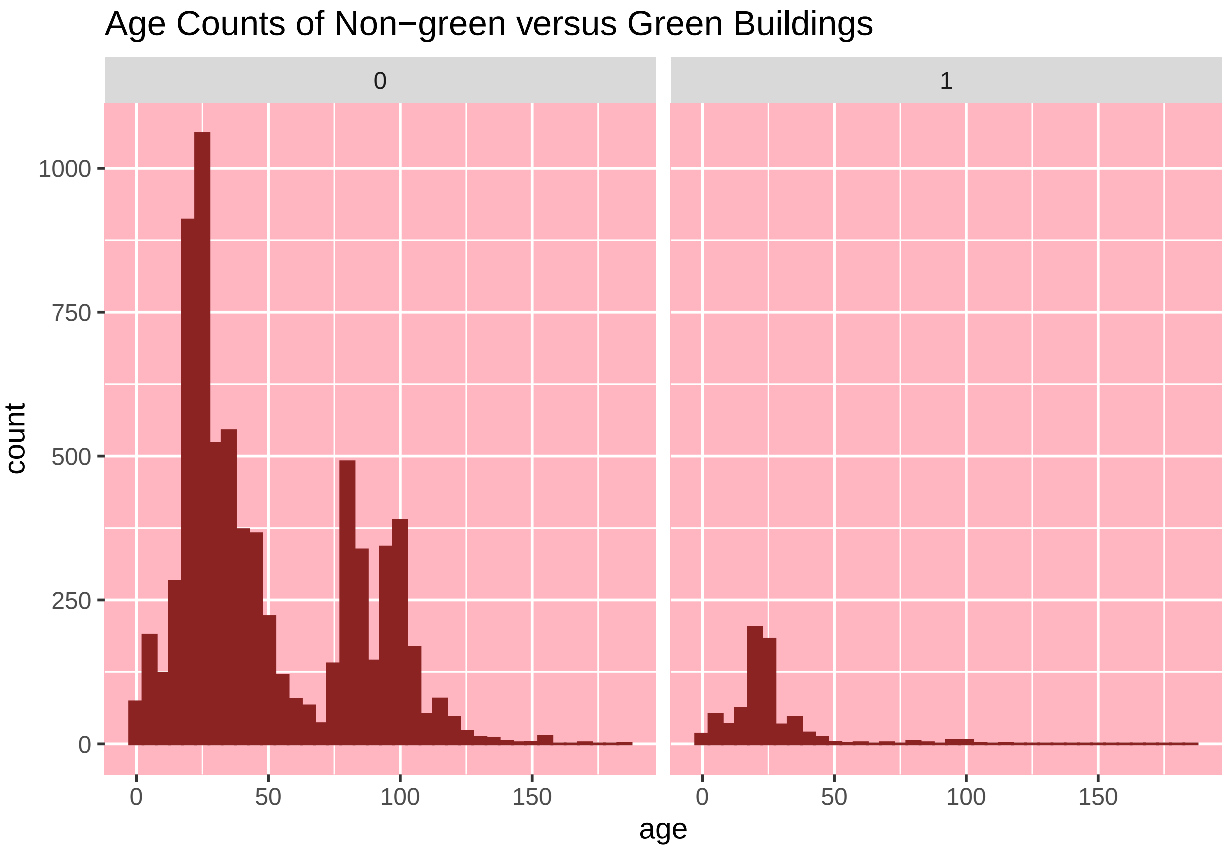
```r
greenb <- read.csv("C:/Users/shigg/OneDrive/Documents/College Notes/R stuff/greenbuildings.txt")

library(tidyverse)
library(mosaic)
library(ggpmisc)


my.formula = y ~ x

ggplot(data = greenb, aes(x = age, y = Rent)) +
  geom_point() +
  ylim(0, 125) +
  geom_smooth(method = "lm") +
  #geom_text(aes(3, 40, , label = paste))+
  geom_smooth(method = "lm", se=FALSE, color="red", formula = my.formula) +
  stat_poly_eq(formula = my.formula, aes(label = paste(..eq.label.., sep = "~~~")), parse = TRUE) +
  theme(panel.background = element_rect(fill = "lightblue", colour = "lightblue", size = 0.5))
```



the slope the line tells us that for every icrease in 1 year of the age of a building, the rent per square foot per year decreases by 0.0527. Since age correllates with rent price, green buildings might not make more profit based solely on rent prices.

now we will see if green buildings are generally younger than non-green buildings

```
numgreen <- (greenb$green_rating == 1)
notgreen <- (greenb$green_rating == 0)

ggplot(data = greenb, aes(x = age)) +
  geom_histogram(binwidth = 5, fill = "brown4", color = "brown4") +
  facet_wrap(~ green_rating, nrow = 1) +
  ggtitle("Age Counts of Non-green versus Green Buildings") +
  theme(panel.background = element_rect(fill = "lightpink", colour = "lightpink", size = 0.5))
```

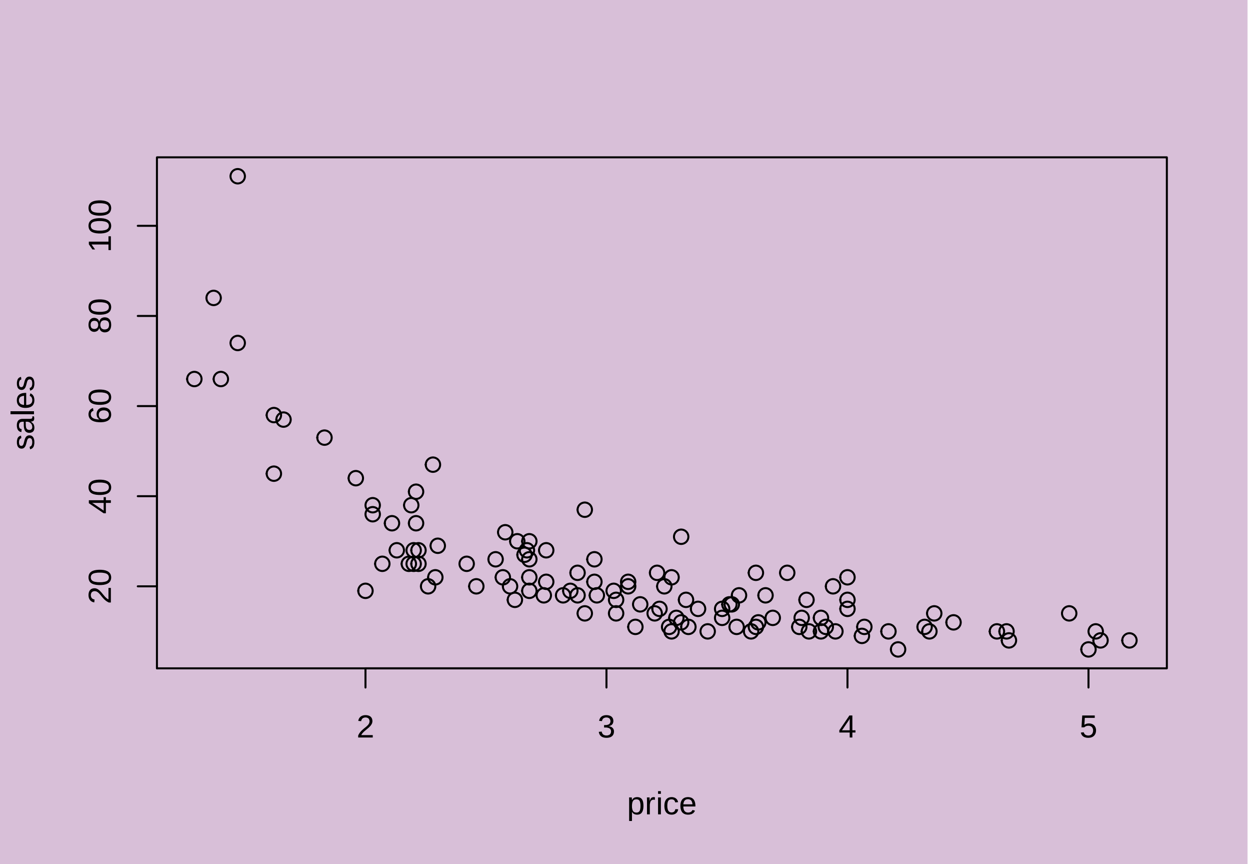


**from this, we can see that the variance of age for nongreen buildings in higher than the variance of age for green buildings. And green buildings proportionally tend to be newer than nongreen buoldings.**

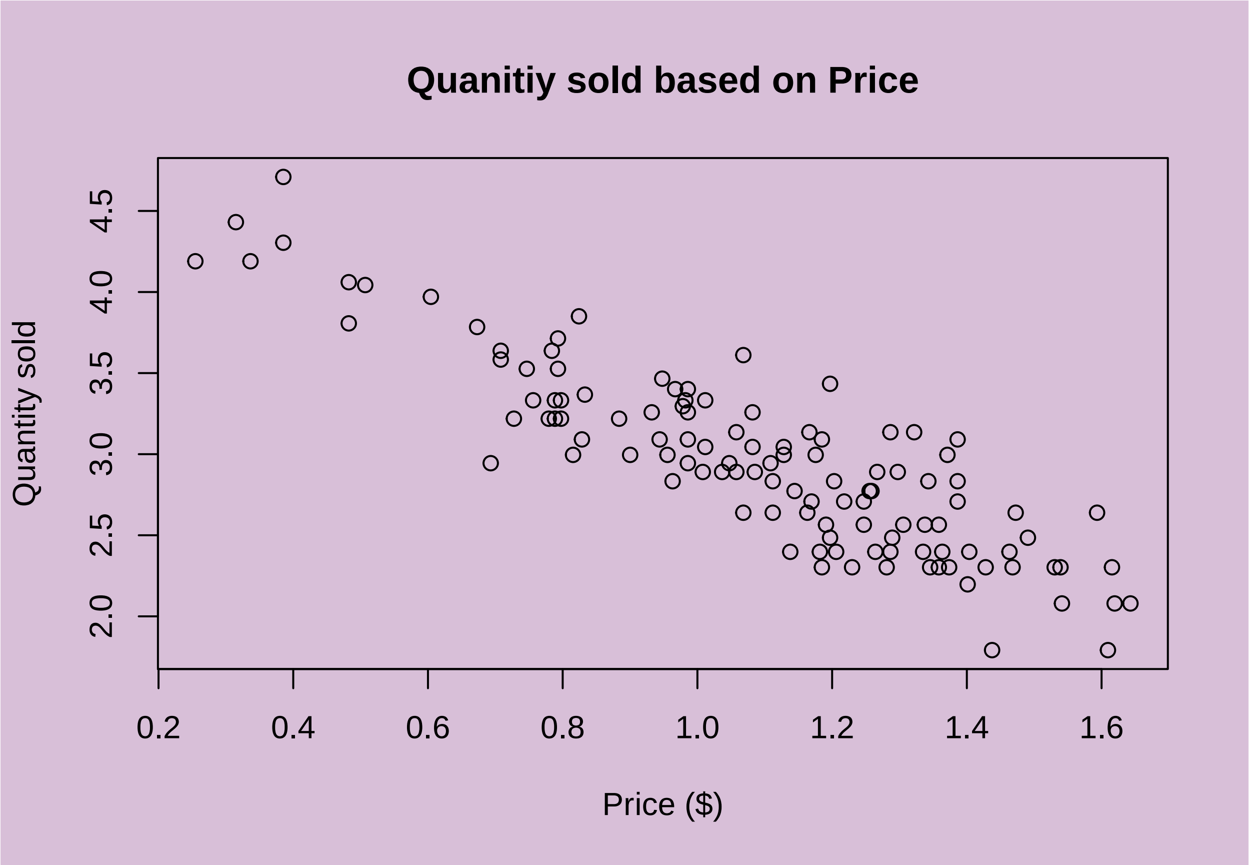
**the analyst's proof behind green buildings having more profit has merit, but he didnt take confounding variables, such as age, that may be a reason for the increased profits in green buildings.**

4. Milk

```
milk <- read.csv("C:/Users/shigg/OneDrive/Documents/College Notes/R stuff/milk.csv")

par(bg = 'thistle')
plot(sales ~ price, data = milk)
```



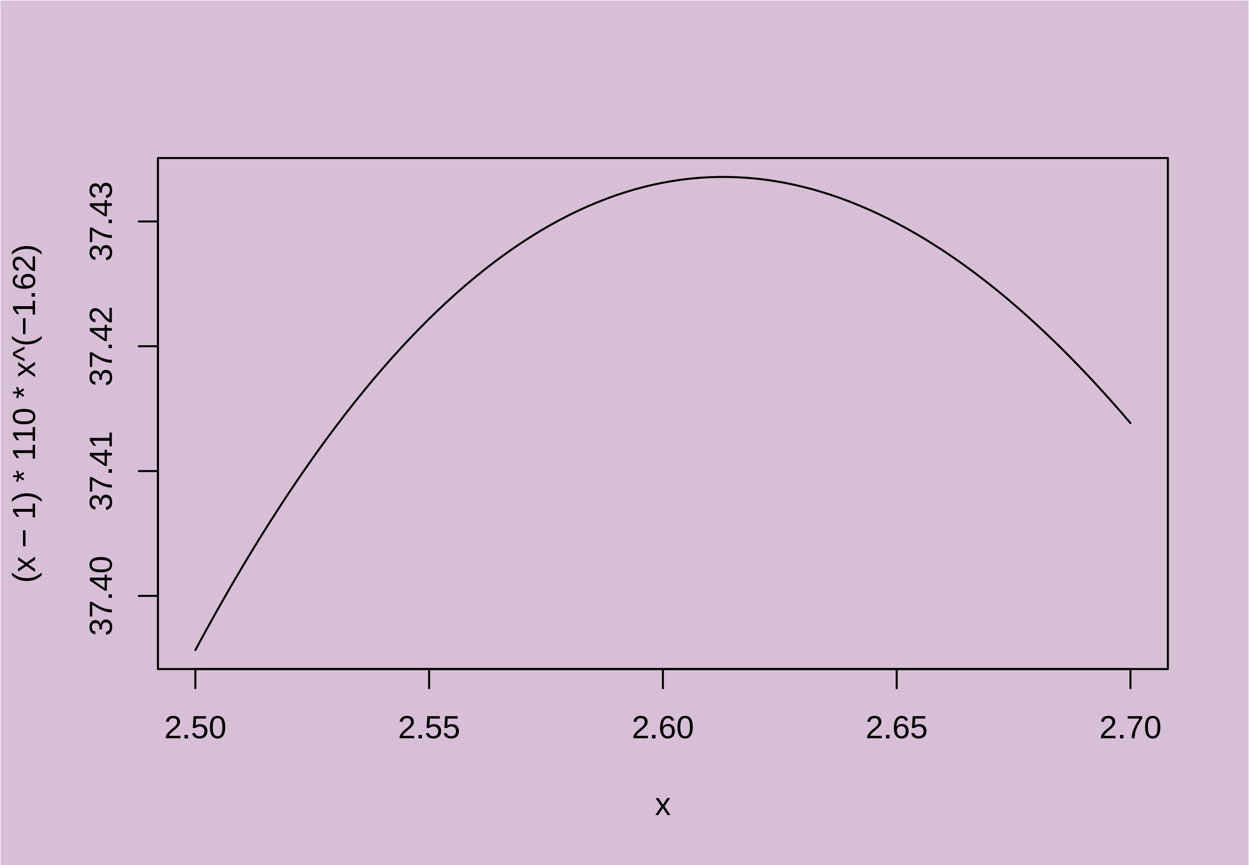
```
plot(log(sales) ~ log(price), data = milk, main = "Quanitiy sold based on Price", xlab = "Price ($)", yl
```



```
lm_ped <- lm(log(sales) ~ log(price), data = milk)
coef(lm)
```

```
## NULL
```

```
curve((x-1)*110*x^(-1.62), from = 2.5, to = 2.7)
```



**The price to get maximum profit is the x point where y is at the maximum point on the curve above, which is approximately between $2.61.**