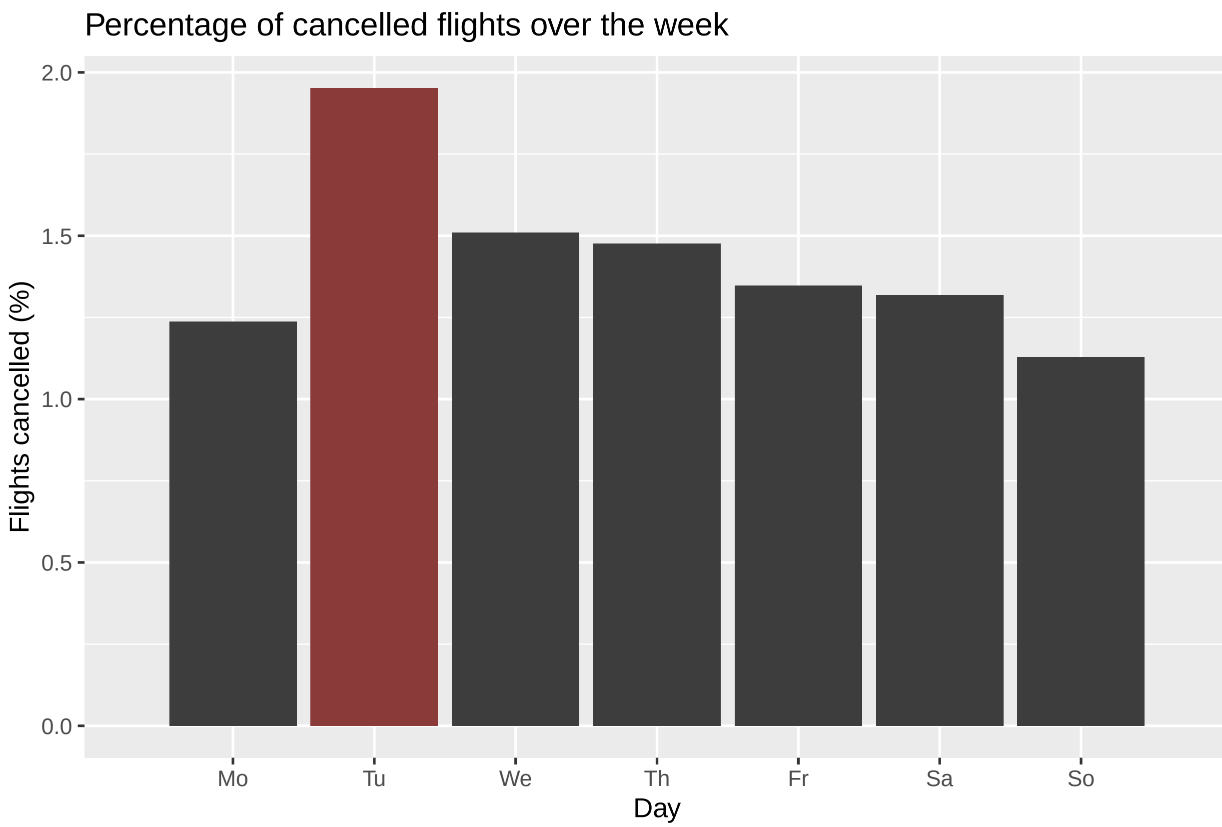
SDS323: Exercise 1

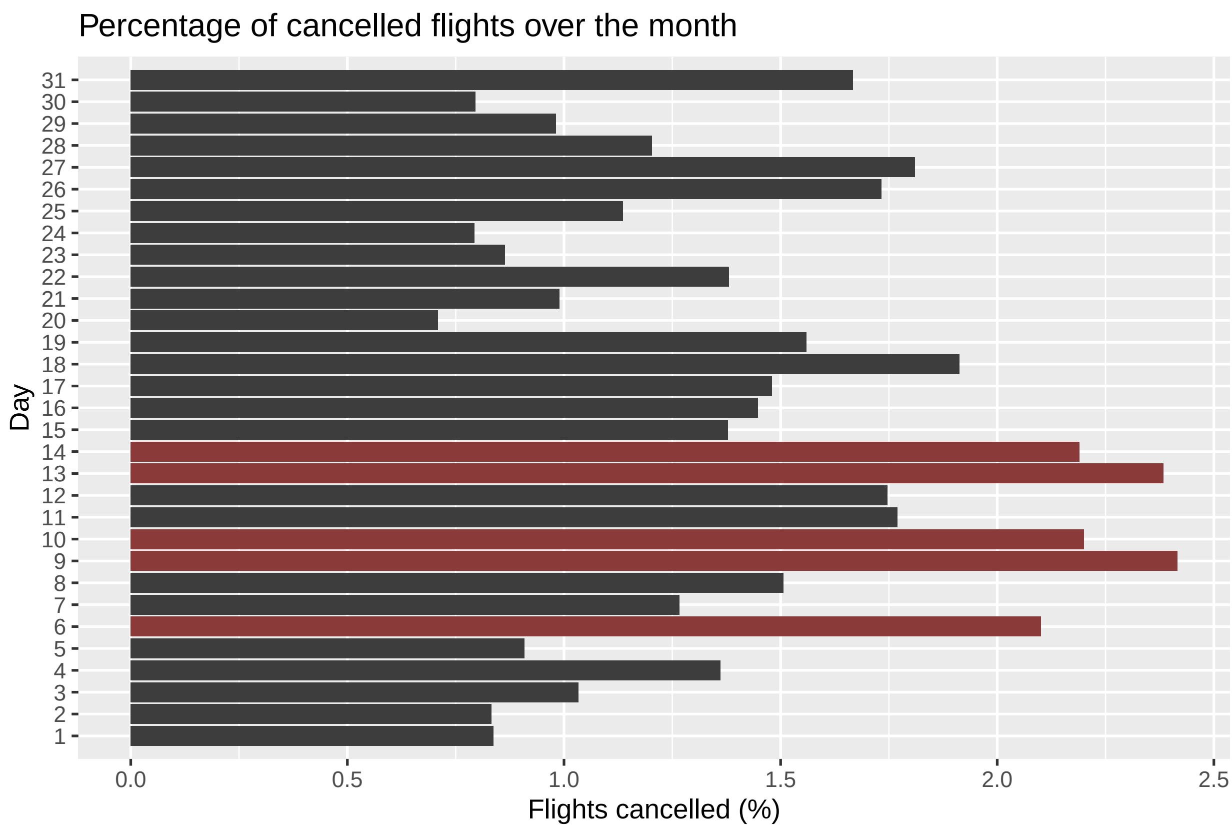
Max Kutschinski (mwk556)

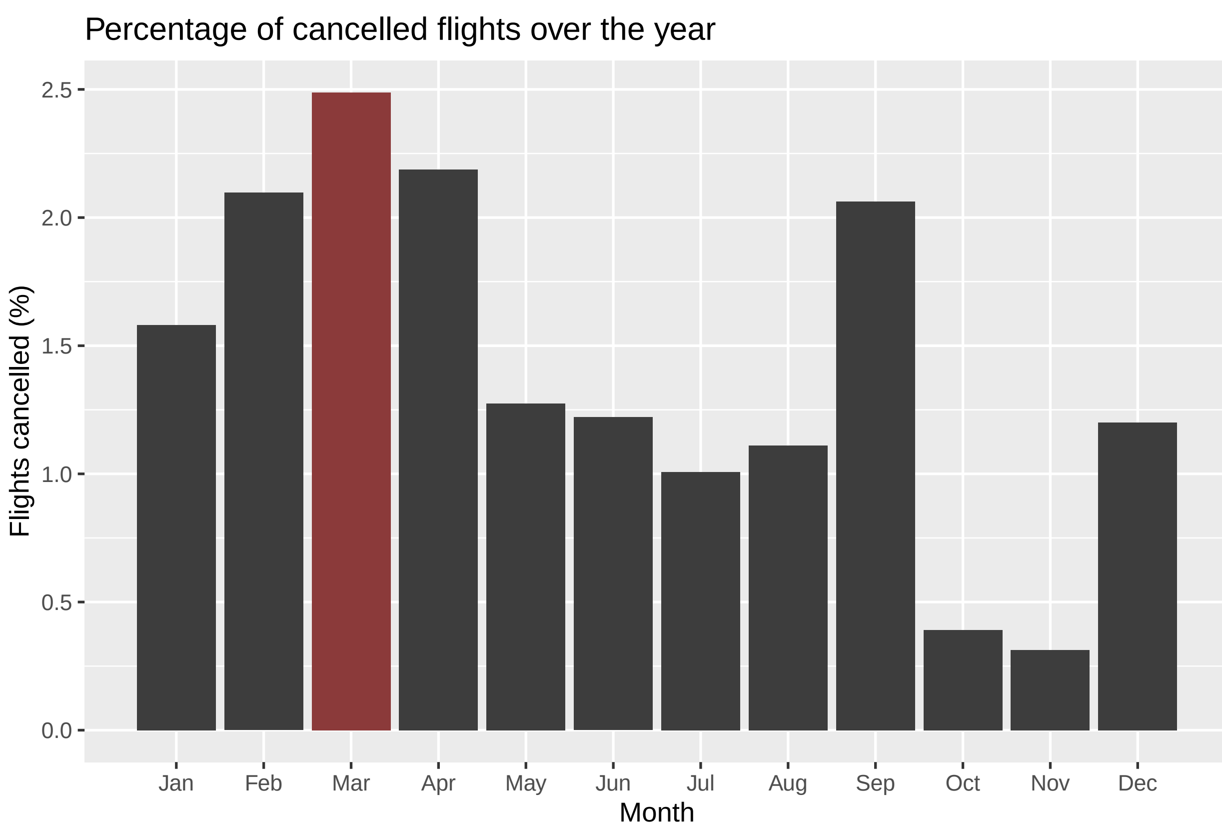
2/14/2020

Data Visualization: Flights at ABIA

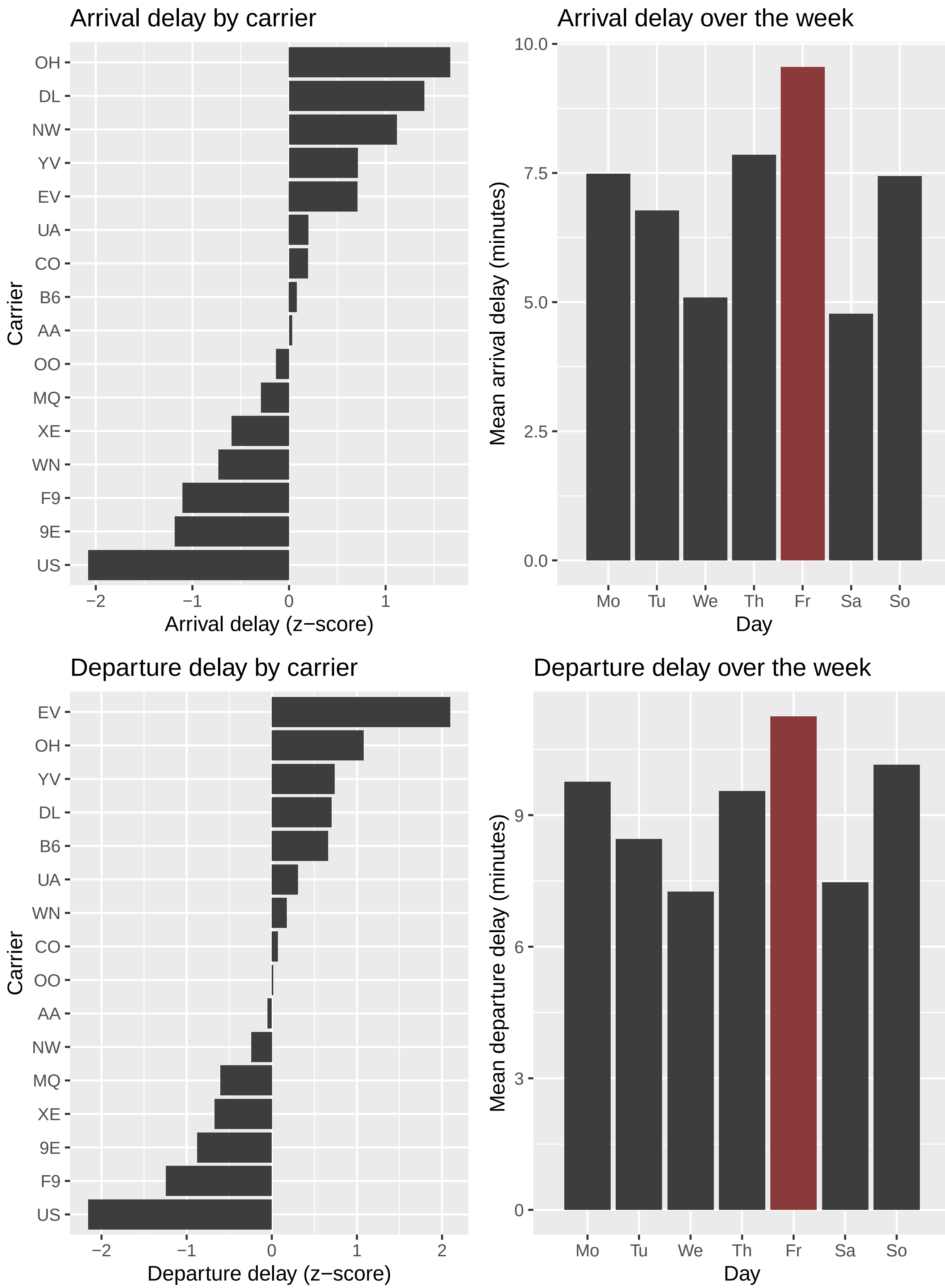
Flight Cancellation Patterns

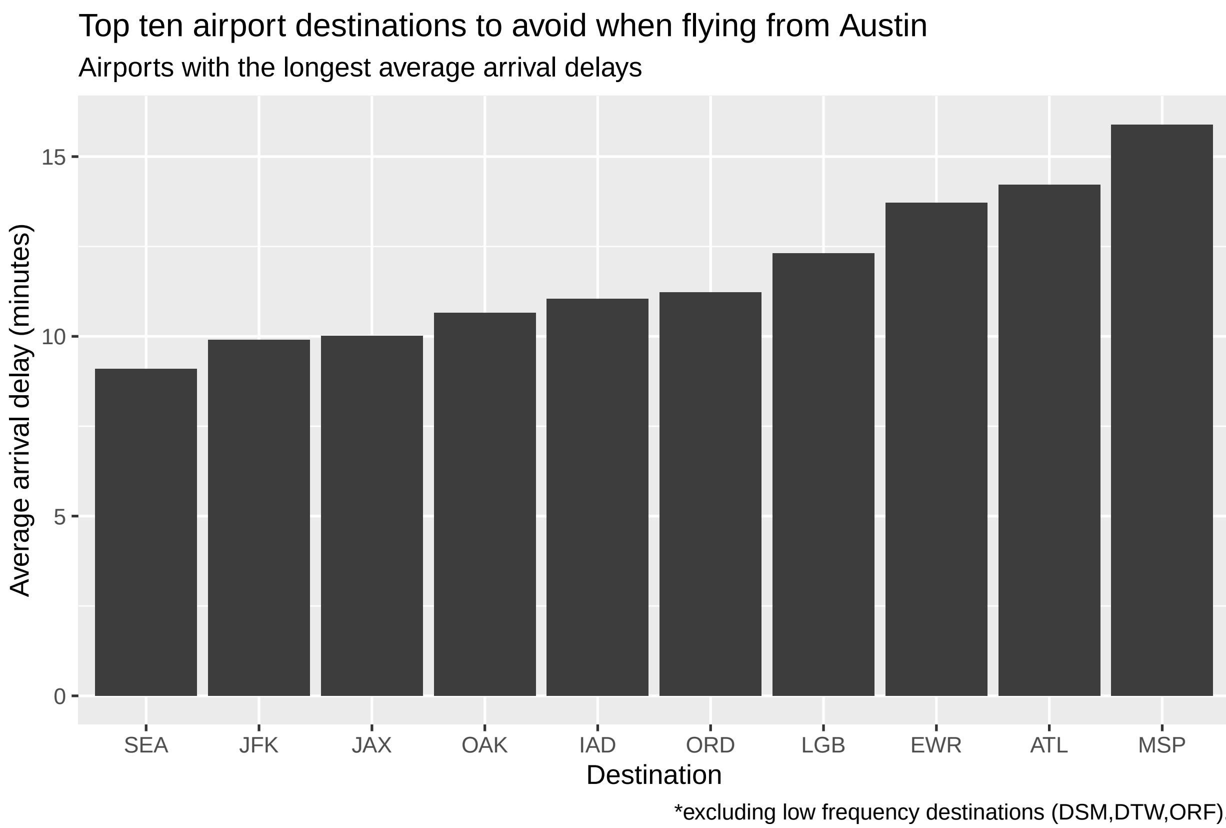



Percentage of cancelled flights over the month
Day
Flights cancelled (%)
0.0
0.5
1.0
1.5
2.0
2.5


Percentage of cancelled flights over the year
Flights cancelled (%)
Month
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec

Flight Delay Patterns




Top ten airport destinations to avoid when flying from Austin
Airports with the longest average arrival delays
Average arrival delay (minutes)
15
10
5
0
SEA
JFK
JAX
OAK
IAD
ORD
LGB
EWR
ATL
MSP
Destination
*excluding low frequency destinations (DSM,DTW,ORF).

Regression Practice

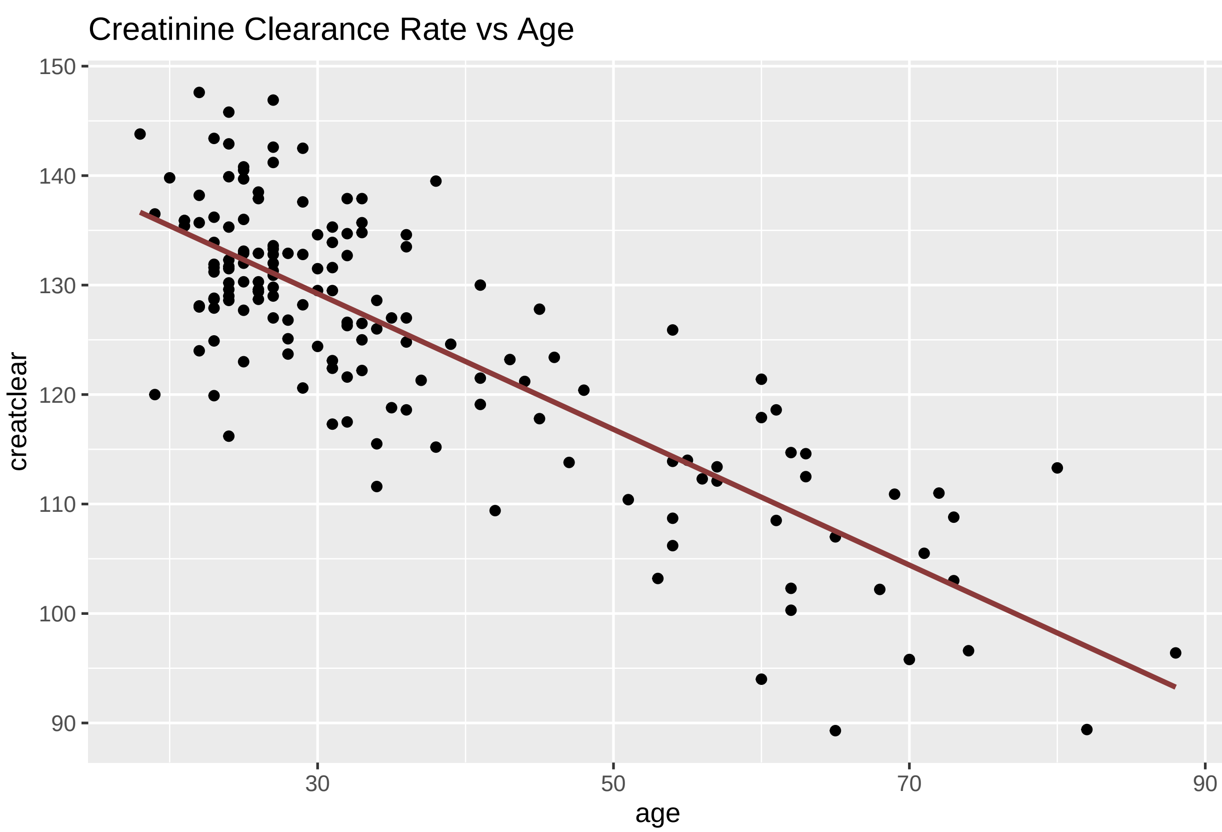
Used creatinine.csv, together with knowledge of linear regression, to answer the following three questions:

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

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

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?

```
ggplot(data=creatinine, aes(age, creatclear)) +
  geom_point() +
  geom_smooth(method="lm", color="indianred4",se= F)+
  labs(title = "Creatinine Clearance Rate vs Age")
```



1. **What creatinine clearance rate should we expect, on average, for a 55-year-old?**

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

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

Therefore, a 55 year old is predicted to have a creatinine clearance rate of 113.723 ml/minute.

2. **How does creatinine clearance rate change with age?**

The rate of change is simply the coefficient of the age variable.

```
coef(lm1)
```

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

Thus, as age increases by one year, clearance rate is predicted to decrease by 0.62 ml/minute. (0.62 ml/minute per year)

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?**

```
pred40=predict(lm1, data.frame(age= 40))
pred60=predict(lm1, data.frame(age= 60))
obs40=135
obs60=112
error40= obs40-pred40
error60= obs60-pred60
error40
```

```
##        1
## 11.97972
```

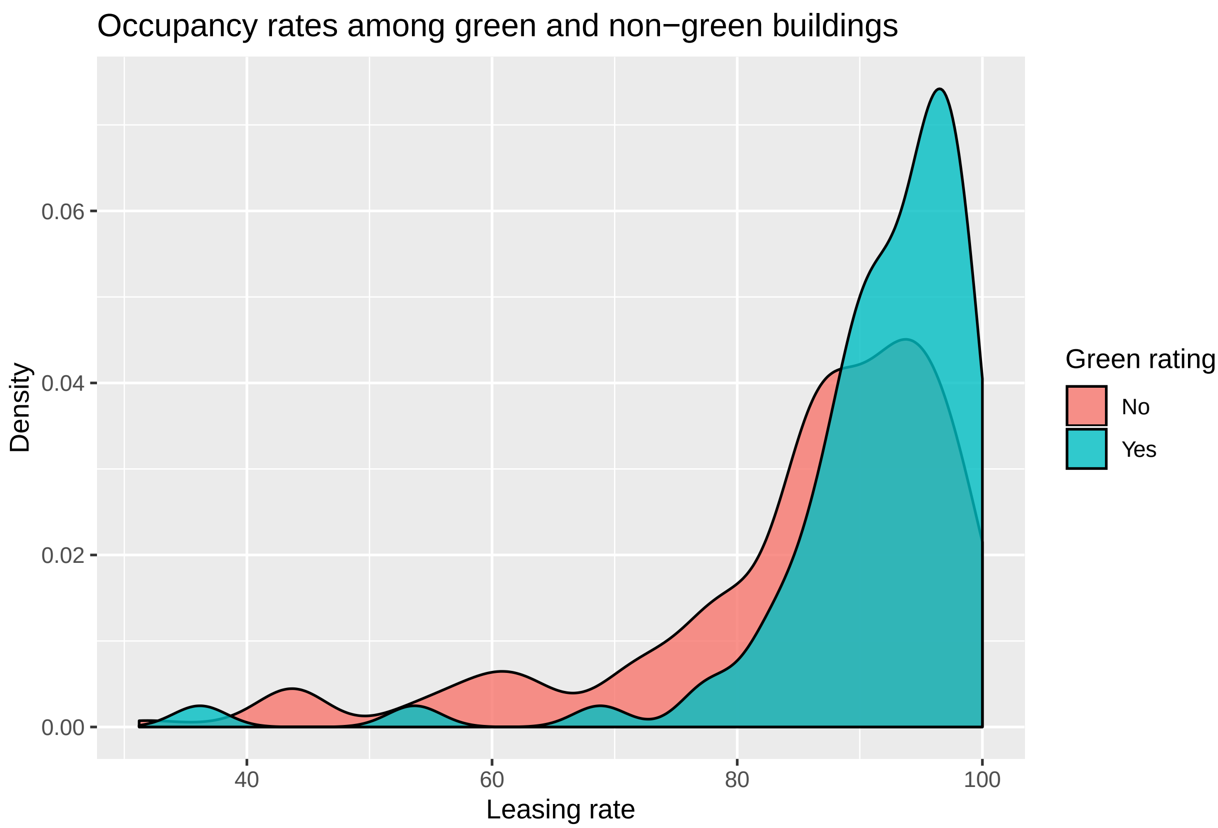
```
error60
```

```
##        1
## 1.376035
```

The 40 year old has a healthier creatine clearance rate for her age, since his observed value is further above the predicted value.

Green Buildings

In order to draw a meaningful conclusion from the dataset, I first cleaned-up it up a little. Clearly, the "data guru" did not control for any confounding variables such as employment growth, utilities, age, etc. In my ananalysis, I decided to restrict the leasing rate to a rate above 30 percent, because I found 10 percent to be a little conservative. Furthermore, I controlled for stories and class, by comparing only good quality buildings with 10 to 20 stories. Lastly, I also restricted the dataset to buildings with a size between 150,000sqft and 350,000sqft in economic areas similar to Austin (employment growth between 2 and 10 percent). Next, I deleted potential outliers that could skew the data.



These two pictures confirm the economic value behind green buildings suggested by the "data guru". While I do not agree with the means he used to obtain his conclusions, I agree for two reasons with the fact that the outlined investment plan is generally a worthwhile investment. First, the rent charged for green buildings is higher on average. Second, the occupancy rates are higher than for similar non-green buildings.

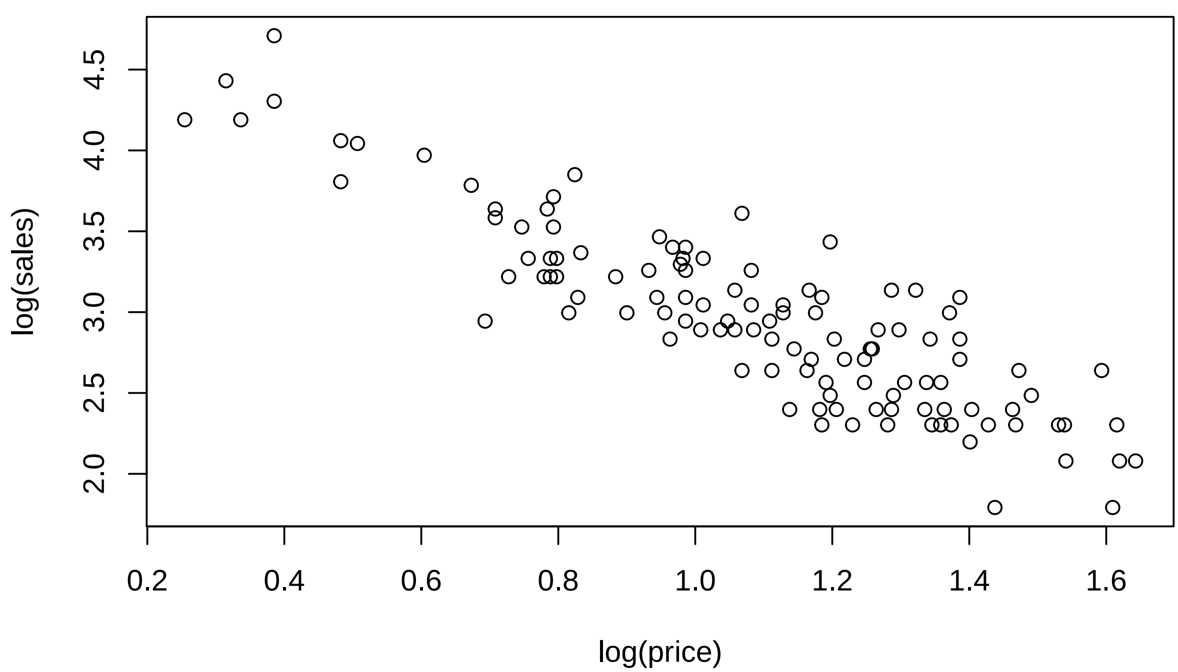
Milk Prices

First, I came up with the profit equation, which depends on the price(P), the quantity(Q), and the cost(C). Note that quantity also depends on price and can therefore be represented as Q(P).

$$Profits = (P - C) * Q(P)$$

Since Q(P) represents the demand curve, I used a log model to make use of the power law. This seems like a good fit, because the log transformation resembles a linear trend.

$$Q(P) = \alpha P^{\beta}$$




```
lm1= lm(log(sales)~log(price), milk)
#coefficients of log model
coef(lm1)
```

```
## (Intercept)  log(price)
##    4.720604   -1.618578
```

```
alpha= exp(coef(lm1)[1])
beta= coef(lm1)[2]
```

The coefficients of this model turned out to be $\alpha \approx 112.24$ and $\beta \approx -1.62$ Thus,

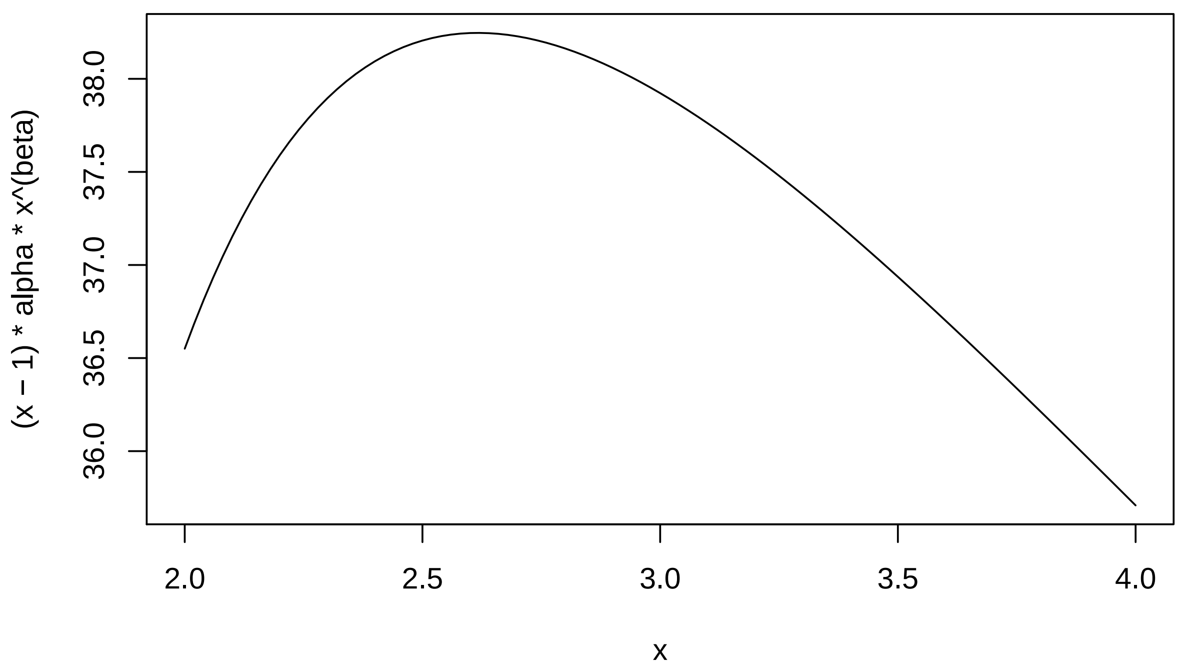
$$Profits = (P - C) * 112.24 * P^{-1.62}$$

Consider Case where C=1. Then the profit maximizing price can be found by setting the first derivate equal to 0 and solving for P:

$$\frac{\partial Profits}{\partial P} = \frac{\partial}{\partial P}[(P - 1) * 112.24 * P^{-1.62}] = 0$$

$$P \approx \$2.61$$

```
curve((x-1)*alpha*x^(beta), from=2, to=4)
```



Given that the cost is equal to \$1, the profit maximizing price is \$2.61. At this price, the net profit is equal to: $Profit = (2.61 - 1) * 112.24 * 2.61^{-1.62} \approx \$38.20$.

In general, the price P* that maximizes the profits for a given per-unit cost C is:

$$\frac{\partial Profits}{\partial P} = \frac{\partial}{\partial P}[(P - C) * 112.24 * P^{-1.62}] = 0$$

$$P^{*} \approx 2.613C$$