SDS 323 Exercise 1

Aaron Grubbs Khue Tran Kaushik Koirala Matthew Tran

02/14/2020

Problem 1: Flights at ABIA

```
## Warning: The labeller API has been updated. Labellers taking `variable`and
## `value` arguments are now deprecated. See labellers documentation.
```

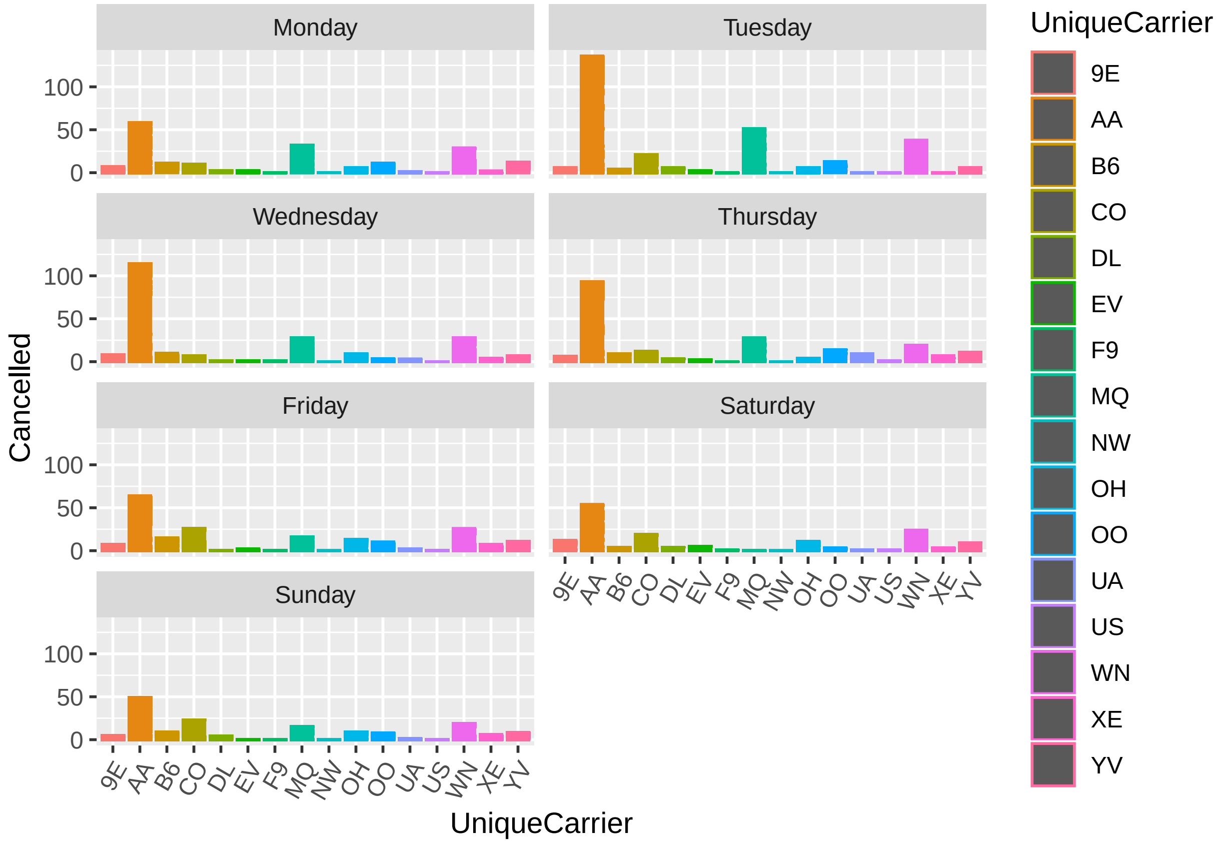
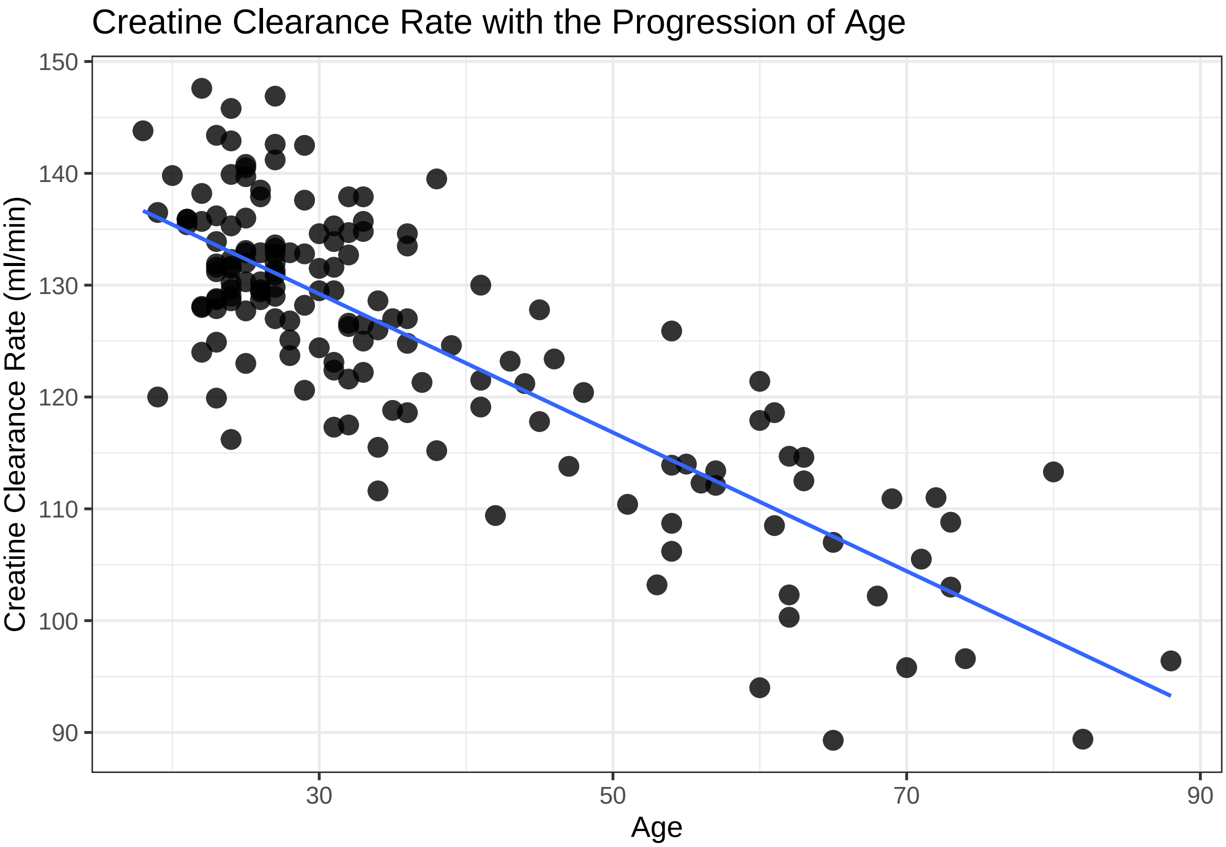


Figure 1: American Airlines flights were the most frequently cancelled, regardless of flight day.

Problem 2: Regression Practice

```
## Parsed with column specification:
## cols(
##   age = col_double(),
##   creatclear = col_double()
## )
```



```
##
## Call:
## lm(formula = creatclear ~ age, data = creatinine)
##
## Coefficients:
## (Intercept)          age
##    147.8129      -0.6198
```

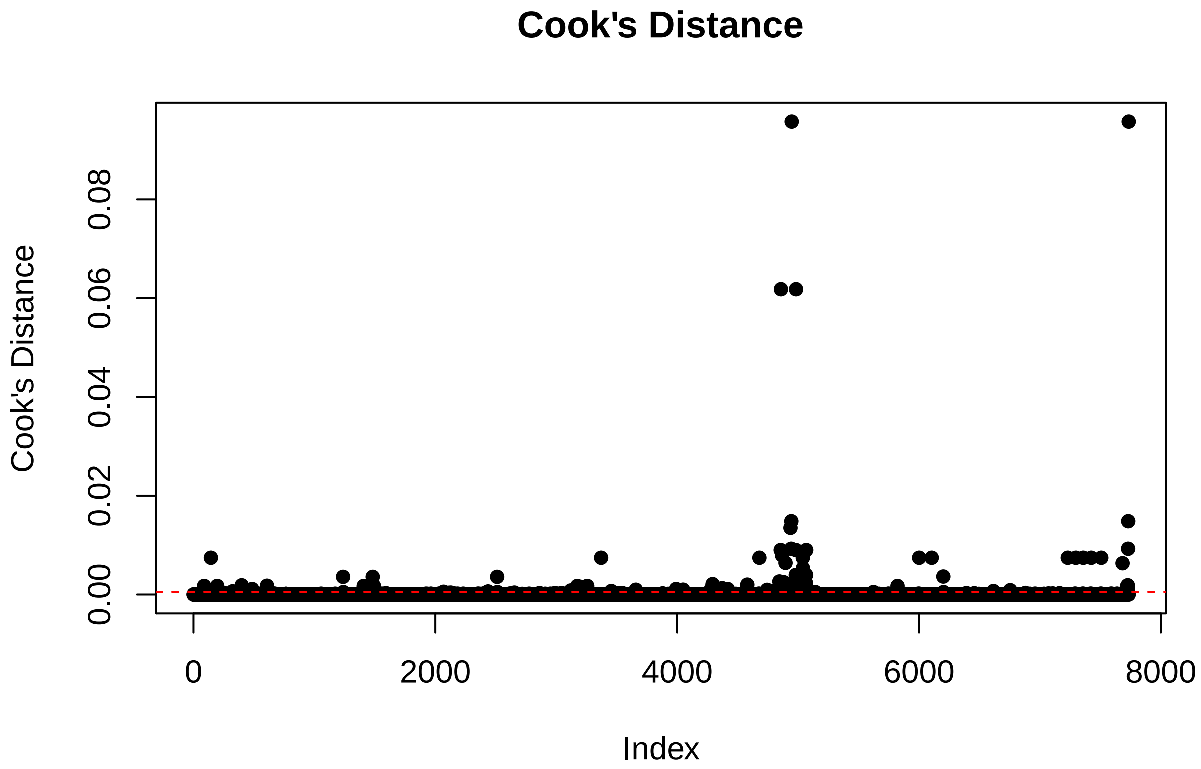
1. We should expect, on average, 113.7239 mL/minute as the creatinine clearance rate for a 55-year old.
2. The slope of -.6198 tells us that for every 1 year increase in age, the average creatinine clearance rate is predicted to decrease by .6198mL/minute.
3. The predicted creatinine clearance rate for a 40 year old is 123.0209 and the predicted creatinine clearance rate for a 60 year old is 110.6249. Because a 40 year old with a creatinine clearance rate of 135mL/min is about 10% above the predicted value (123.209mL/min) by the regression line for a 60 year old and because the creatinine clearance rate of 112mL/minute is above the predicted value (110.6249mL/minute) only by about 1 percent, it is healthier to be a 40 year old with a 135mL/min creatinine clearance rate than a 60 year old with a 112mL/min creatinine clearance rate, assuming that the spreads of the creatinine clearance rates at each of those ages are reasonably similar.

Problem 3: Green Buildings

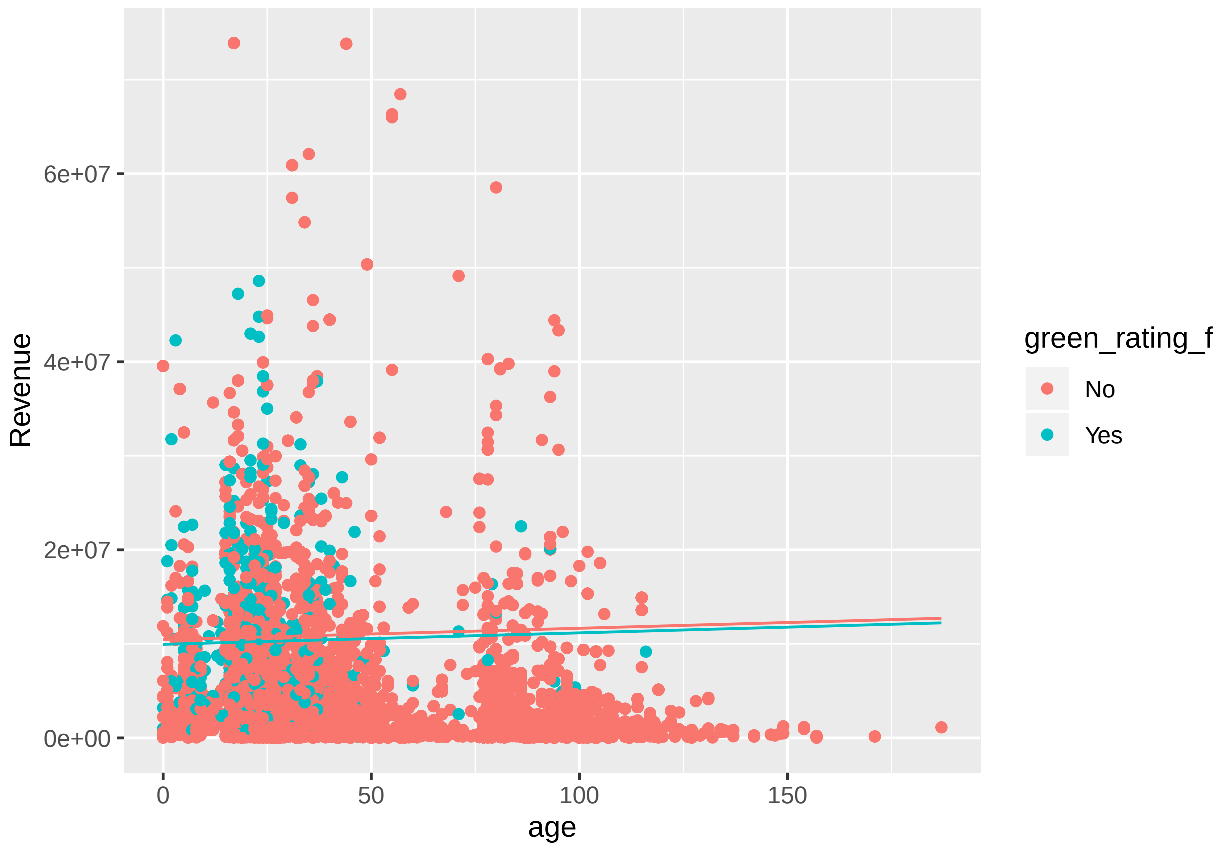
```
## Parsed with column specification:
## cols(
##   .default = col_double()
## )
## See spec(...) for full column specifications.

##
## Call:
## lm(formula = Revenue ~ green_rating_f + amenities_f + age + renovated_f +
##     class_col_f, data = green)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -14310664  -4471620  -1213919   1078289 455597151
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -1907857     666594  -2.862  0.00422 **
## green_rating_fYes -1754368     616967  -2.844  0.00447 **
## amenities_fYes     3611140     372937   9.683  < 2e-16 ***
## age                  37488       7029   5.333 9.93e-08 ***
## renovated_fYes     -997200     397753  -2.507  0.01219 *
## class_col_fB       2118578     542199   3.907 9.41e-05 ***
## class_col_fA      10914394     644600  16.932  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14810000 on 7727 degrees of freedom
## Multiple R-squared:  0.1038, Adjusted R-squared:  0.1031
## F-statistic: 149.2 on 6 and 7727 DF,  p-value: < 2.2e-16

##                    GVIF Df GVIF^(1/(2*Df))
## green_rating_f 1.081533  1        1.039968
## amenities_f    1.219148  1        1.104150
## age            1.804126  1        1.343178
## renovated_f    1.316070  1        1.147201
## class_col_f    1.657228  2        1.134607
```

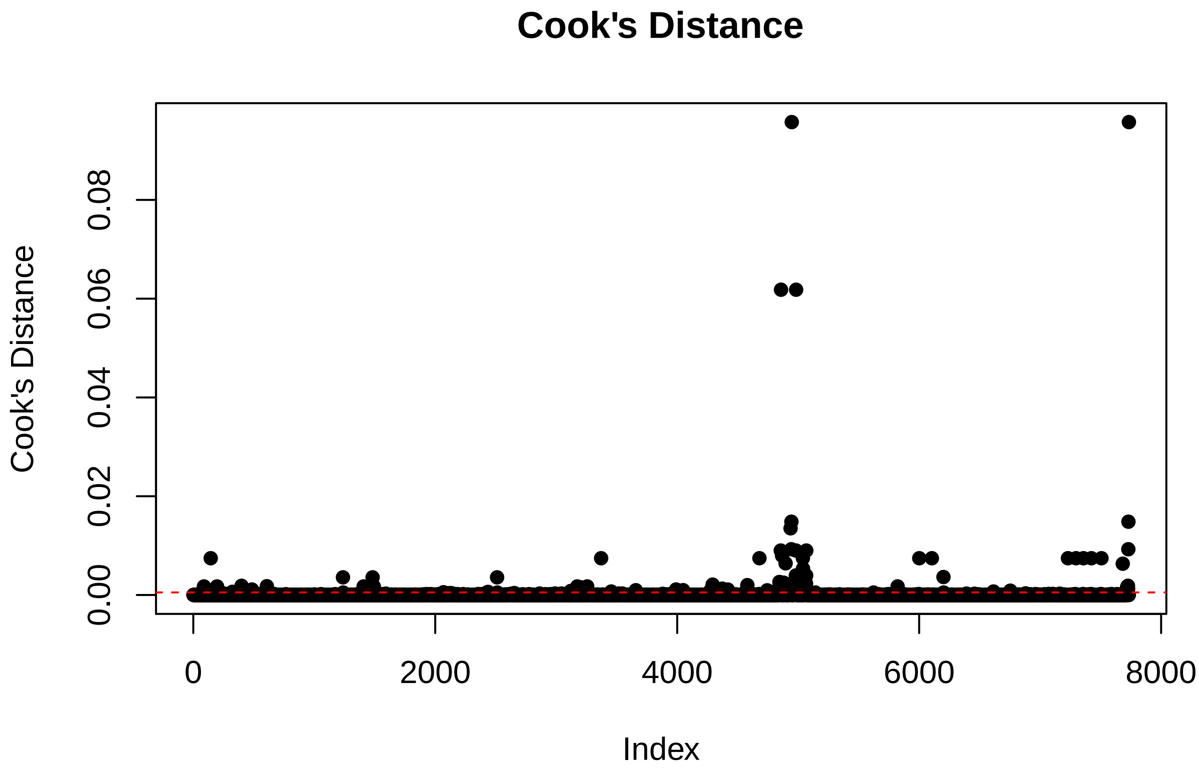


```
##
## Call:
## lm(formula = Revenue ~ green_rating_f + amenities_f + age + renovated_f +
##     class_col_f, data = green_g)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -10750552  -3198693   -916964    870118  63182612
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -274264     293810  -0.933   0.3506
## green_rating_fYes  -490203     272553  -1.799   0.0721 .
## amenities_fYes     2841266     164601  17.262  < 2e-16 ***
## age                  12189       3115   3.913 9.21e-05 ***
## renovated_fYes      -65510     176062  -0.372   0.7098
## class_col_fB       1544035     238431   6.476 1.00e-10 ***
## class_col_fA       7944378     284835  27.891  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6510000 on 7658 degrees of freedom
## Multiple R-squared:  0.2666, Adjusted R-squared:  0.2661
## F-statistic: 464.1 on 6 and 7658 DF,  p-value: < 2.2e-16
```

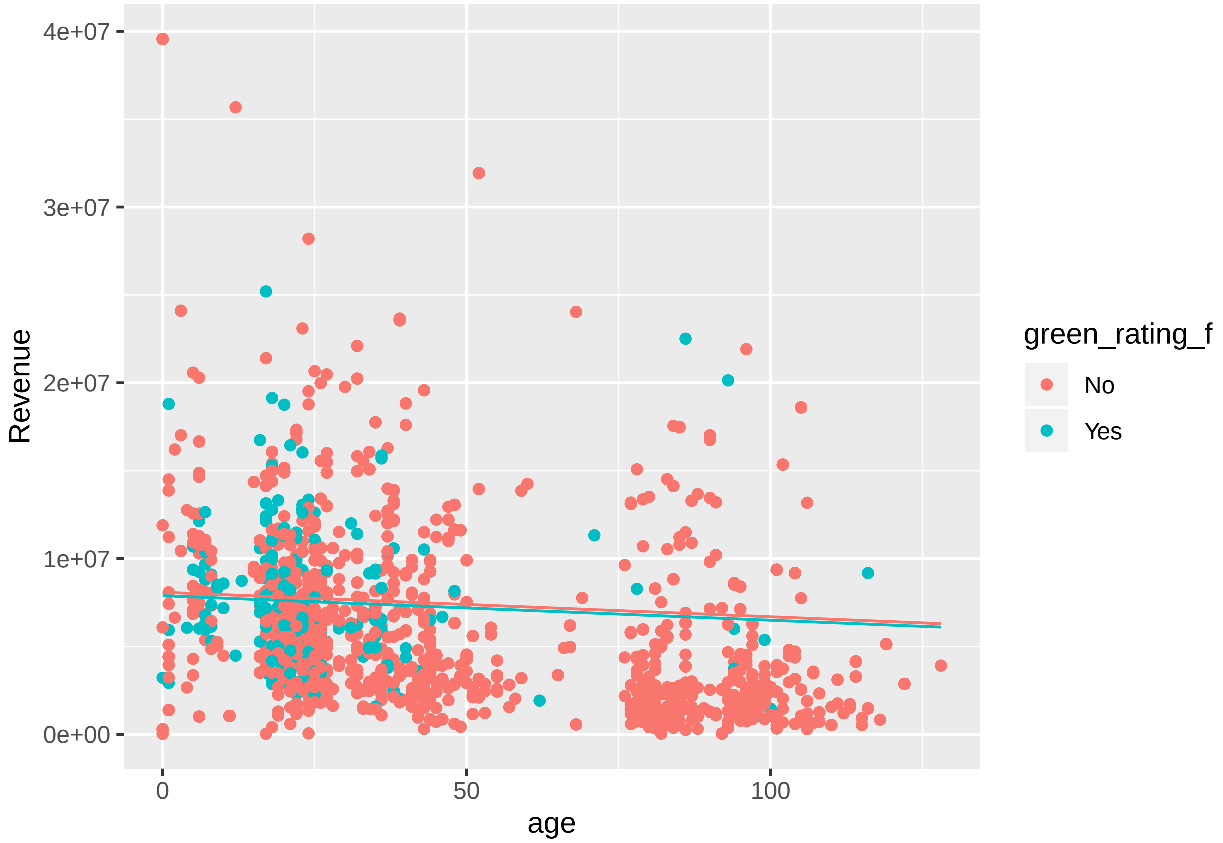


```
##
## Call:
## lm(formula = Revenue ~ green_rating_f + amenities_f + age + renovated_f +
##     class_col_f, data = green15)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8597866 -2459474  -933379  1278221 30570933
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3434477     429509   7.996 1.97e-15 ***
## green_rating_fYes  -196086     302063  -0.649 0.516299
## amenities_fYes     1381010     186799   7.393 1.97e-13 ***
## age                 -13961       3812  -3.663 0.000255 ***
## renovated_fYes     -901421     197025  -4.575 5.00e-06 ***
## class_col_fB       1396561     335063   4.168 3.18e-05 ***
## class_col_fA       4170825     375599  11.104  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4114000 on 2412 degrees of freedom
## Multiple R-squared:  0.2333, Adjusted R-squared:  0.2314
## F-statistic: 122.3 on 6 and 2412 DF,  p-value: < 2.2e-16
```

```
##                    GVIF Df GVIF^(1/(2*Df))
## green_rating_f 1.095686  1        1.046750
## amenities_f    1.129103  1        1.062593
## age            2.045321  1        1.430147
## renovated_f    1.364025  1        1.167915
## class_col_f    1.629617  2        1.129852
```



```
##
## Call:
## lm(formula = Revenue ~ green_rating_f + amenities_f + age + renovated_f +
##     class_col_f, data = green15_g)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8597866 -2459474  -933379  1278221 30570933
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        3434477     429509   7.996 1.97e-15 ***
## green_rating_fYes  -196086     302063  -0.649 0.516299
## amenities_fYes     1381010     186799   7.393 1.97e-13 ***
## age                 -13961       3812  -3.663 0.000255 ***
## renovated_fYes     -901421     197025  -4.575 5.00e-06 ***
## class_col_fB       1396561     335063   4.168 3.18e-05 ***
## class_col_fA       4170825     375599  11.104  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4114000 on 2412 degrees of freedom
## Multiple R-squared:  0.2333, Adjusted R-squared:  0.2314
## F-statistic: 122.3 on 6 and 2412 DF,  p-value: < 2.2e-16
```

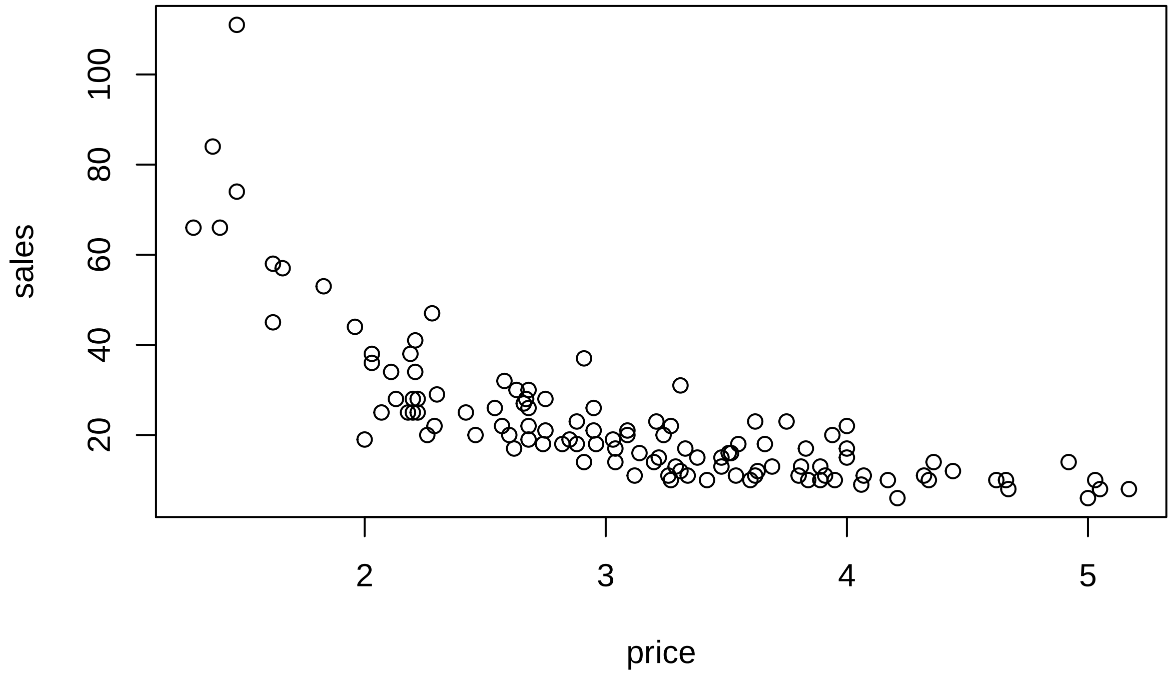


Right off the bat the guru does not attempt to consider other variables that would affect the rent and thus the projected revenue. Controlling for these other variables gives a larger picture on the actual revenues these buildings are receiving. Comparing the revenues between buildings could prove to be beneficial since it should consider both the rent and occupancy. As it is now there is an assumption that the rent given by the guru is definitely the one that will be used but in actuality it will likely change depending on other factors that again the guru did not account for. The timeline to profitability is also off since it is unknown how the occupancy rate will change once the building is on the market which would affect the breakeven point.

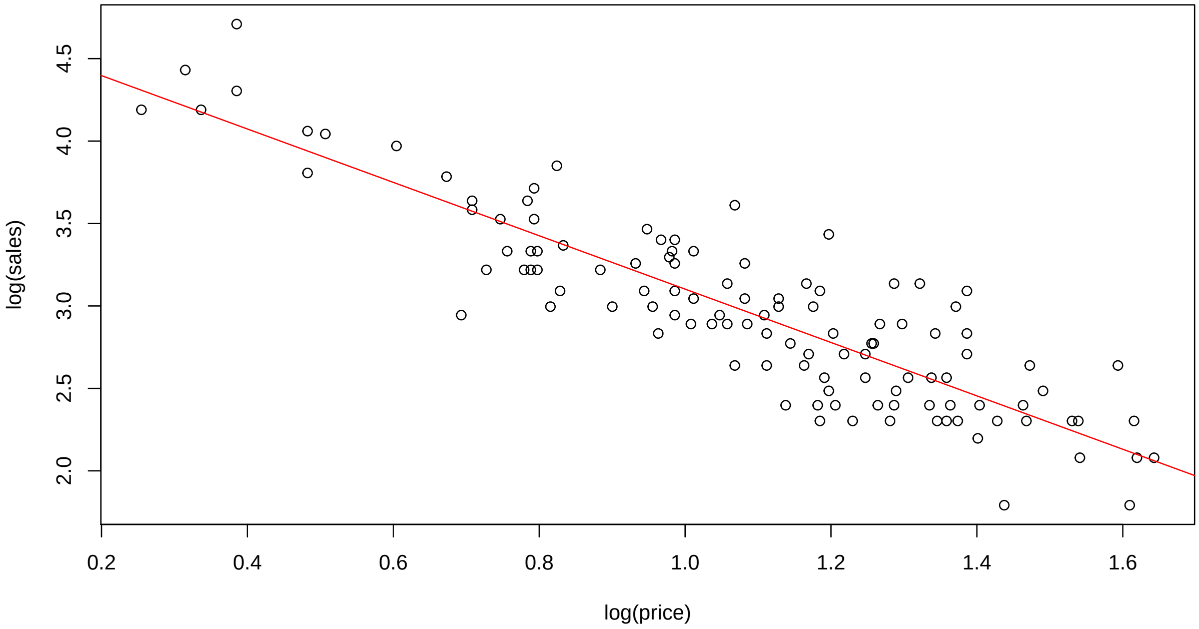
The regression provided is based on the rough estimation of the revenue based on size, occupancy, and rent. Based on the data provided it does not matter whether or not the building has a green certification or not since it does not have a significant impact on income based on the summary of the regression model. The model includes variables that are more easily controlled by the builders, such as amenities and building quality or can be reasonably accounted for such as age. Variables such as utilities and climate were not included due to the complication of how each building pays for utilities and the uncertainty of how the climate may or may not change over time. A second regression was run to filter the buildings that had roughly the same number of stories as the building being proposed (+/-5 stories) since that is known. This regression keeps the same conclusion as the previous one. Of note the other variables, building quality and amenities, do have a significant positive impact on the income.

Problem 4: Milk

The following plot is the plot of quantity of sales vs price.

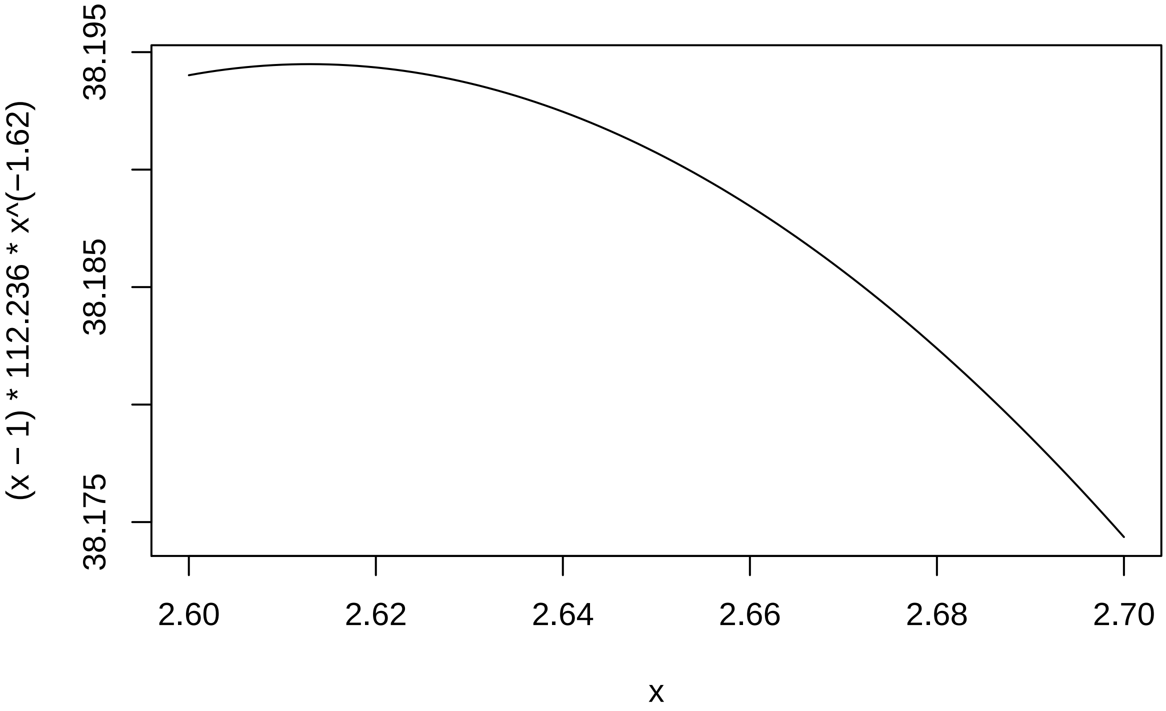


Because sales price and quantity sold are coupled, we have to use the power law. More generally, we have to get the coefficients of the regression line between log(sales) and log(price). This is what that looks like:



The coefficients for this regression are 4.7206042 and -1.6185778.

Using the first value above as the coefficient of β_0, $\alpha = e^{\beta_0}$ or $\alpha = 112.2360465$. With this we get the following equation:
Profit = (price-1) * 112.2360465 * $price^{4.7206042}$
The figure plotted out, where the x-axis is the price, and the y-axis is the profit, looks like this:



The curve shows us that the max profit is around a price point of 2.61. Using a little bit of calculus we can verify that:

```
## Loading required package: mosaicCore

##
## Attaching package: 'mosaicCore'

## The following object is masked from 'package:plyr':
##
##     count

## The following object is masked from 'package:car':
##
##     logit

## The following objects are masked from 'package:dplyr':
##
##     count, tally

##
## Attaching package: 'mosaicCalc'

## The following object is masked from 'package:stats':
```

```
##
##     D

## function (x)
## 112.236 * x^(-1.62) + (x - 1) * 112.236 * (x^((-1.62) - 1) *
##     (-1.62))

## $root
## [1] 2.612915
##
## $f.root
## [1] -6.383488e-05
##
## $iter
## [1] 7
##
## $init.it
## [1] NA
##
## $estim.prec
## [1] 6.103516e-05
```

The only practical root is indeed at price point $2.61, giving us a max profit of $38.1944653