SDS Exercise 1

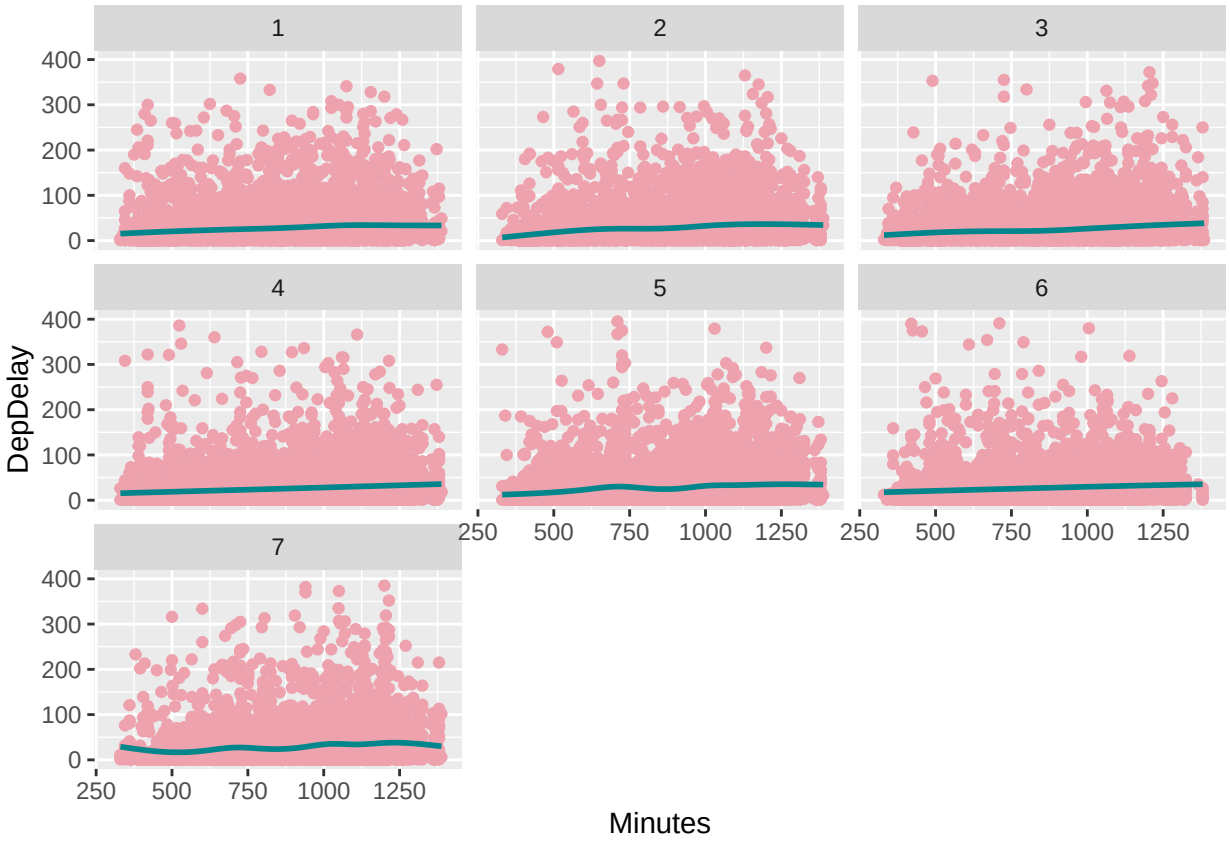
ABIA

We assessed whether the data supports a conclusion that there are times of the day that are systematically preferable to other times of the day from the standpoint of minimizing delays in departures. ABIA commences scheduled departures at 5:30 am, and the data indicates that flights at that time typically depart a few minutes ahead of schedule, and that on-time departures gradually deteriorate over the course of the day. This is reflected in the scatter plot below, in which the data for Departure Delays has been filtered and limited to those flights with a Departure Delay > 0. On average, flights after approximately 6:00 depart some minutes after their scheduled departure time and by evening the delays are in excess of 20 or more minutes. This pattern holds generally true on each day of the week, and on each month of the year. The experience with delayed arrivals is broadly similar:

```
library(mosaic)
library(tidyverse)
```

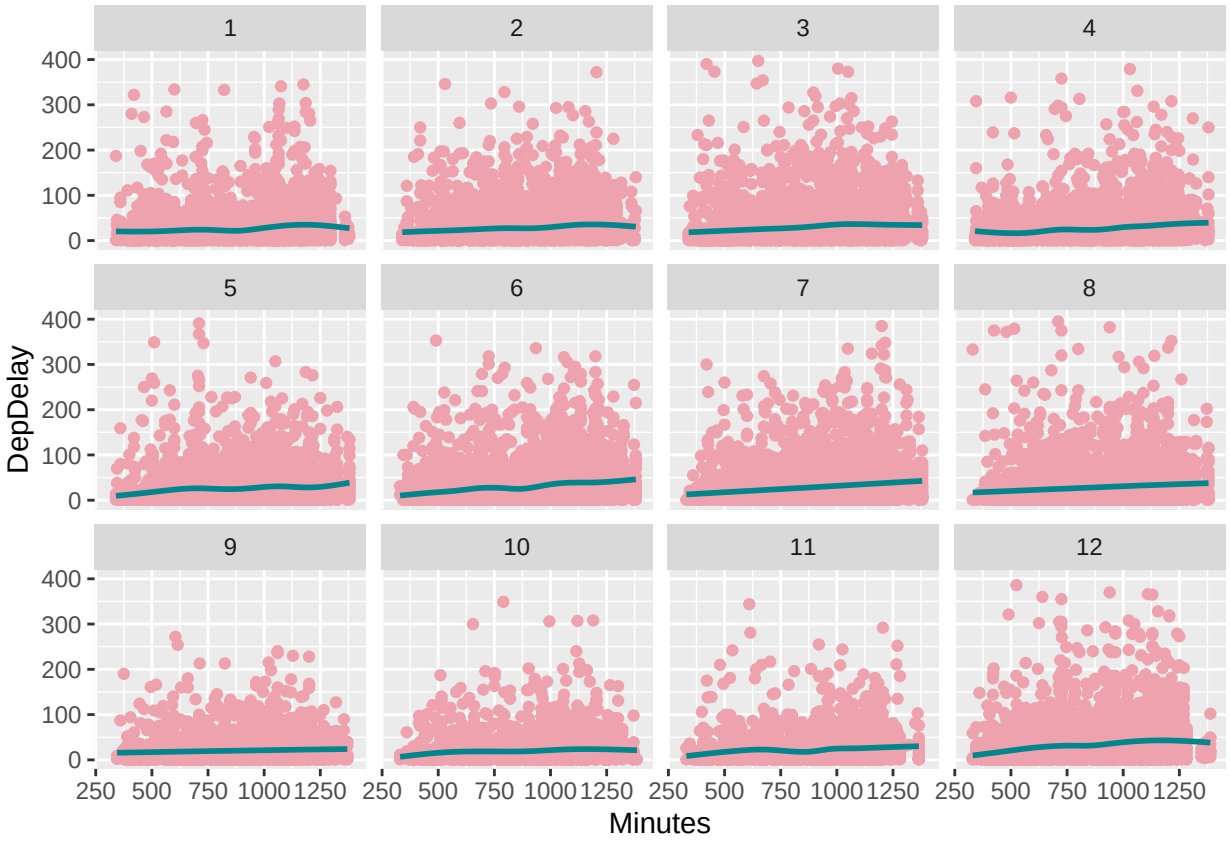
Delay & Day of the week

```
ABIA <- read.csv("/Users/pengcheng/Desktop/UT Austin/Spring 2020/SDS 323/Exercise 1/ABIA.csv")
ABIA=filter(ABIA, DepDelay>0)
ABIA=mutate(ABIA, Minutes=CRSDepTime%/%100 * 60 + CRSDepTime %% 100)
sp<-ggplot(data=ABIA)+
  geom_point(mapping=aes(x=Minutes,y=DepDelay),color = 'lightpink2')+
  geom_smooth(mapping=aes(Minutes,DepDelay),se = FALSE, color = 'turquoise4')+
  facet_wrap(~DayOfWeek)
sp+xlim(300,1400)+ylim(0,400)
```



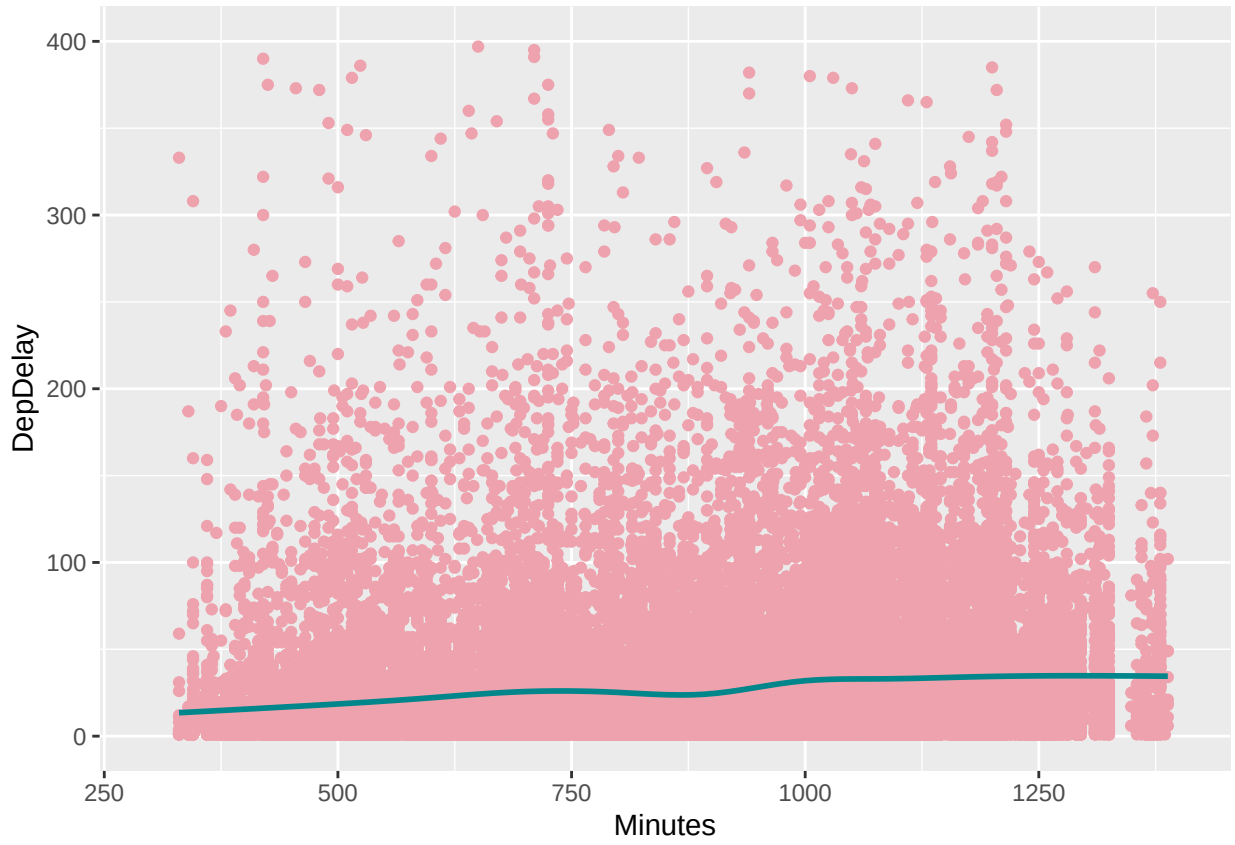
Delay & Month of the year

```
sp<-ggplot(data=ABIA)+
  geom_point(mapping=aes(x=Minutes,y=DepDelay), color = 'lightpink2')+
  geom_smooth(mapping=aes(Minutes,DepDelay),se = FALSE, color = 'turquoise4')+
  facet_wrap(~Month)
sp+xlim(300,1400)+ylim(0,400)
```



Overall trend

```
sp<-ggplot(data=ABIA)+
  geom_point(mapping=aes(x=Minutes,y=DepDelay), color = 'lightpink2')+
  geom_smooth(mapping=aes(Minutes,DepDelay),se = FALSE, color = 'turquoise4')
sp+xlim(300,1400)+ylim(0,400)
```

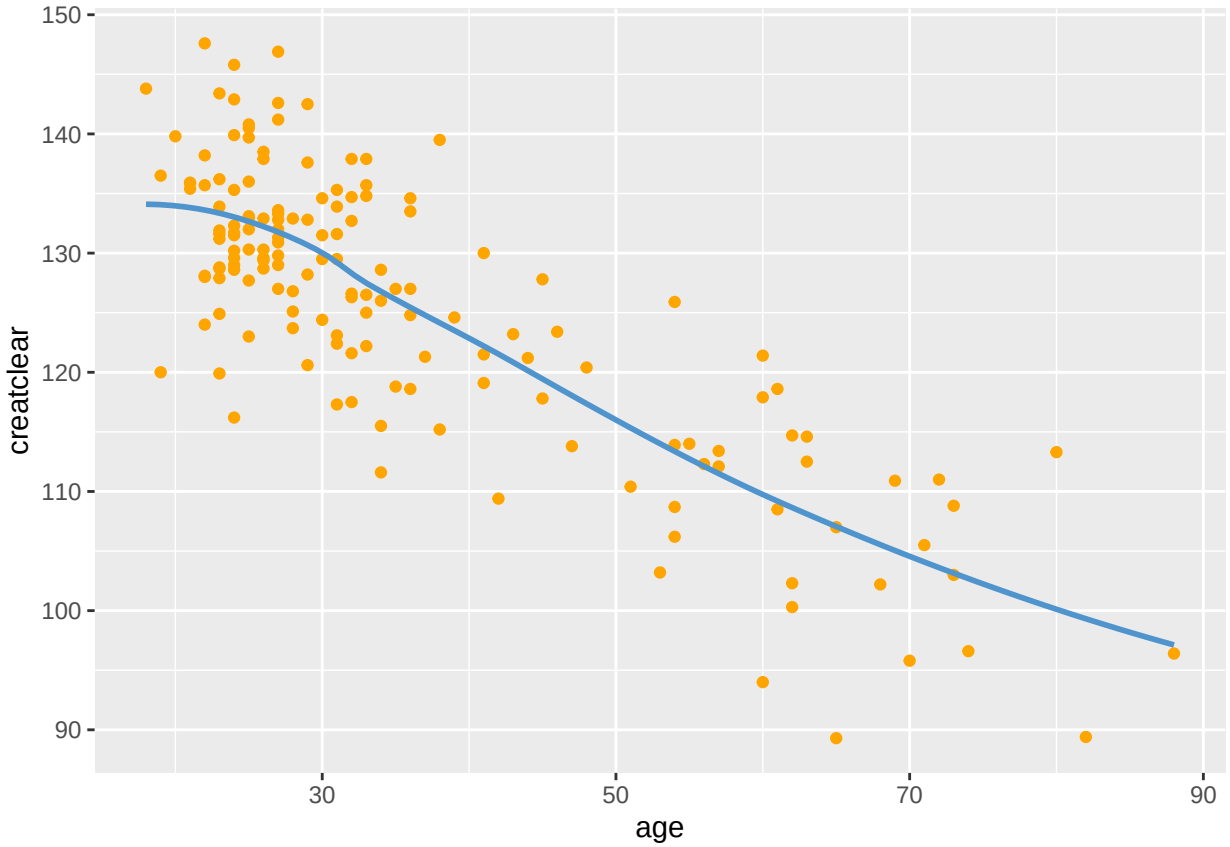


(minutes are running minutes during the day, such that 300 is 5:00am)

Regression

We prepared a scatter plot reflecting the Creatinine dataset, and plotted a geom_smooth line based on that data:

```
creatinine <- read.csv("/Users/pengcheng/Desktop/UT Austin/Spring 2020/SDS 323/Exercise 1/creatinine.csv
ggplot(data=creatinine)+
  geom_point(mapping=aes(x=age,y=creatclear), color = 'orange')+
  geom_smooth(mapping=aes(x=age,y=creatclear),se = FALSE, color = 'steelblue3')
```




```
lm_creat = lm(creatclear ~ age, data = creatinine)
coef(lm_creat)
```

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

Therefore, based on the available data we conclude that the f(x) for creatinine clearance rate is as follows:

creatclear=-0.6198159*age+147.8129158

Questions:

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

- A 55 year old patient would have an expected creatinine clearance rate of 113.723 ml/minute.

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

- Based on the data, we anticipate an individual's creatinine clearance rate to decline with age as follows: creatinine clearance rate(age)=-0.6198159ml/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?

- The expected creatinine clearance rate for a 40 year-old=123.0203 ml/minute, which is 11.9797 lower than our test subject.
- The expected creatinine clearance rate for a 60 year-old=110.624 ml/minute, which is 1.376 lower than our test subject.
- Although both subjects are healthier than the average for their age, the 40 year old is healthier than the 60 year old on both an absolute and age adjusted basis.

Green Building

```
greenbuildings <- read.csv("/Users/pengcheng/Desktop/UT Austin/Spring 2020/SDS 323/Exercise 1/greenbuil
```

```
greenbuilding = filter(greenbuildings, leasing_rate > 10)
```

We first assume the methodology employed by the analyst to be correct, which is to filter out all buildings with a leasing rate smaller than 10. Under this consumption, we generate the new data set greenbuilding. The following analysis will be based on this new data set greenbuilding.

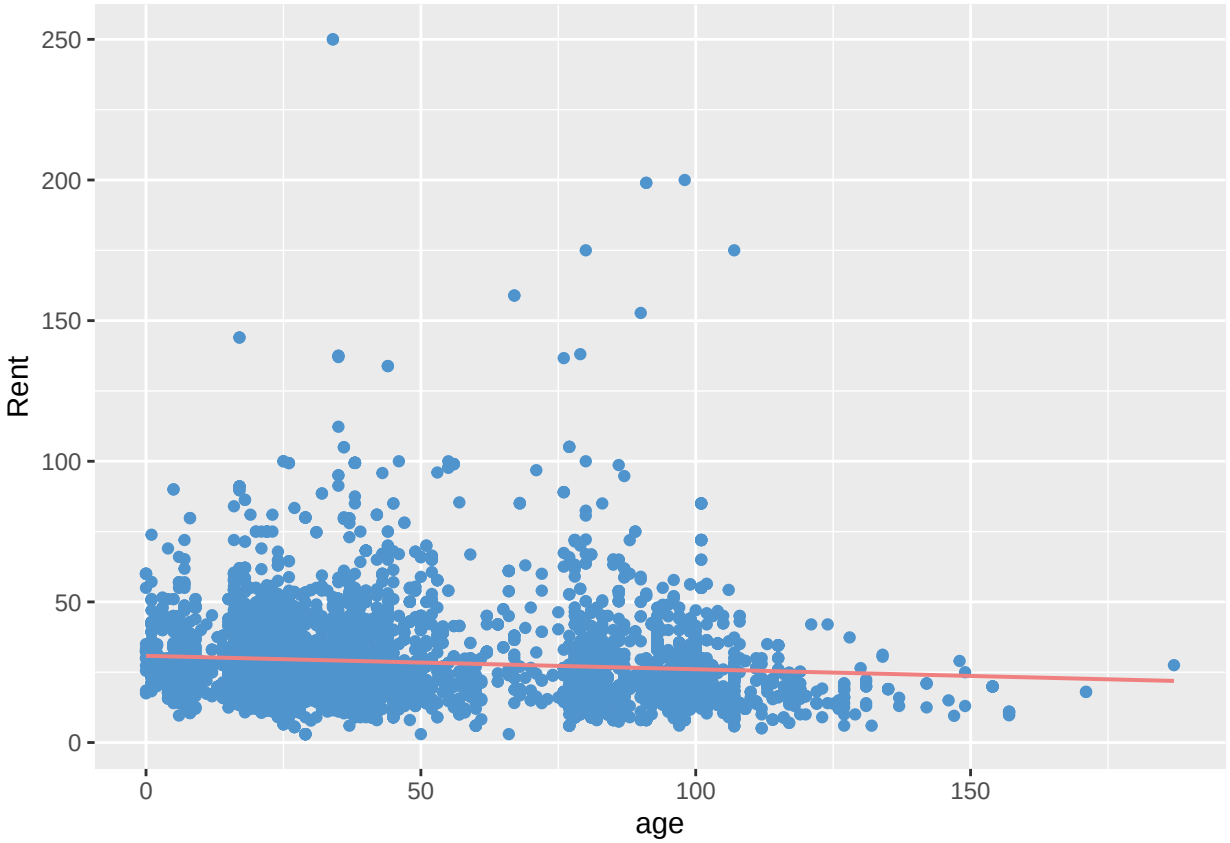
To look at the profitability of buildings, we examine the effect green certification has on rents and the occupancy rates.

Rent

The rent of buildings is contingent on many factors like amenities, age of the building, etc. Though on average, green buildings do seem to have a higher rent than non-green buildings, it could very well be the result of, for instance, green buildings are newer. Therefore, to estimate the effect of green certification on the rent, we have to examine the correlation between green_rating and confounding variables like age.

Rent & Age

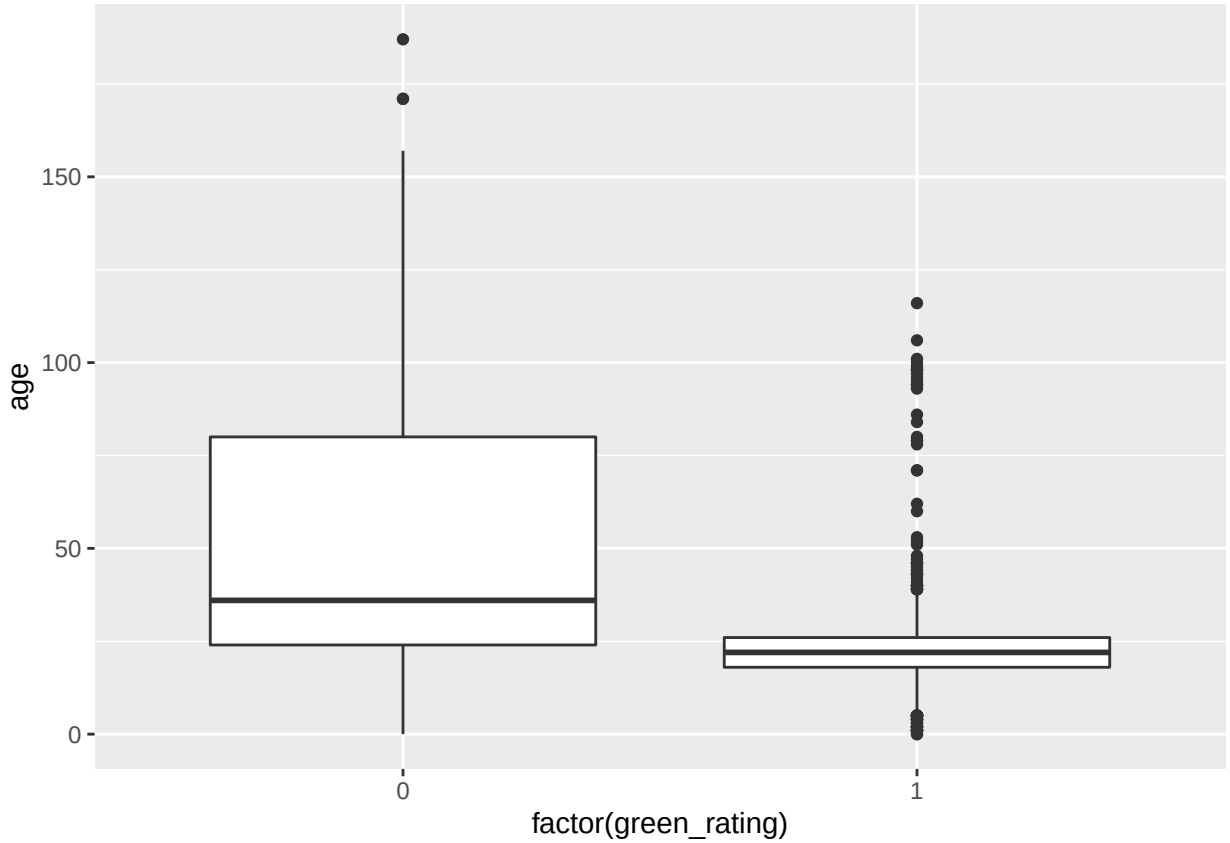
```
ggplot(data = greenbuilding) +
  geom_point(mapping = aes(x = age, y = Rent), color = 'steelblue3') +
  geom_lm(mapping = aes(x = age, y = Rent), color = 'lightcoral')
```



The above scatter plot shows the relationship between rent and age. As the age of the building increases, the rent of the building tends to decrease. This is consistent with our common perceptions, which is that newer buildings tend to have higher rents than the old ones.

Next, we investigate the relationship between green_rating and age.

```
ggplot(data = greenbuilding) +
  geom_boxplot(mapping=aes(x=factor(green_rating), y=age))
```



The box plot gives a visualization of the age distribution for green & non-green buildings. Clearly, buildings with green certifications are generally newer than non-green buildings.

```
lm1 = lm(age ~ green_rating, data=greenbuilding)
coefficients(summary(lm1))
```

```
##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)    49.30808  0.3746663 131.60533 0.000000e+00
## green_rating -25.42796  1.2553617 -20.25549 6.494996e-89
```

The regression analysis presents the same picture. According to the regression model, green buildings are typically 25.43 years newer than non-green buildings. Our previous findings show that newer buildings do tend to have a higher rent. Namely, part of the reason green buildings, on average, have a higher rent than non-green buildings could be that green buildings are typically newer.

Rent & Class

Next, we investigate the relationship between rents and class.

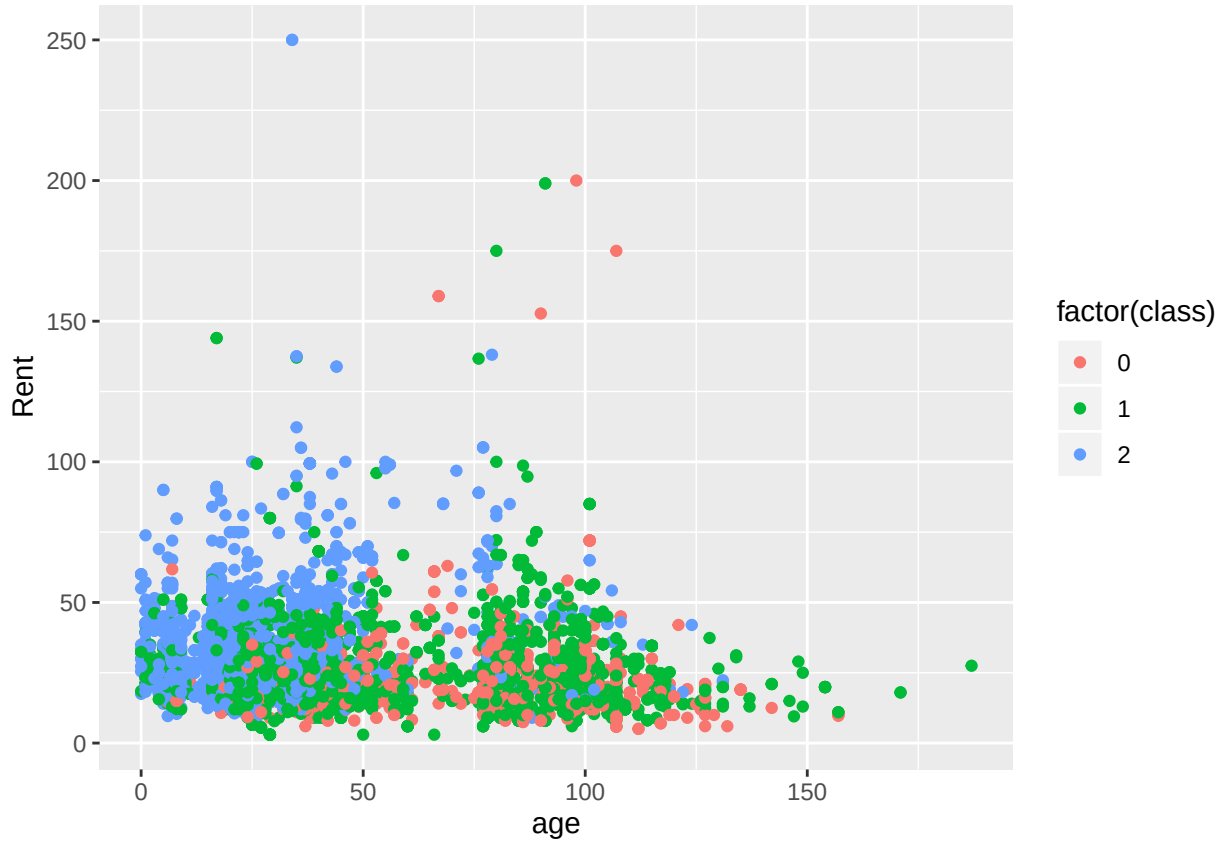
To better represent the class of each building, we generate a new variable class. Buildings of class_a have value 2, buildings of class_b have value 1, and buildings of class_c have value 0.

```
greenbuilding = mutate(greenbuilding, class = class_a * 2 + class_b * 1)
```

Then, as we believe buildings of higher quality may be buildings built in relatively recent years. Thus, age and class of buildings might be correlated. So, in order to better investigate the relationship between Rent

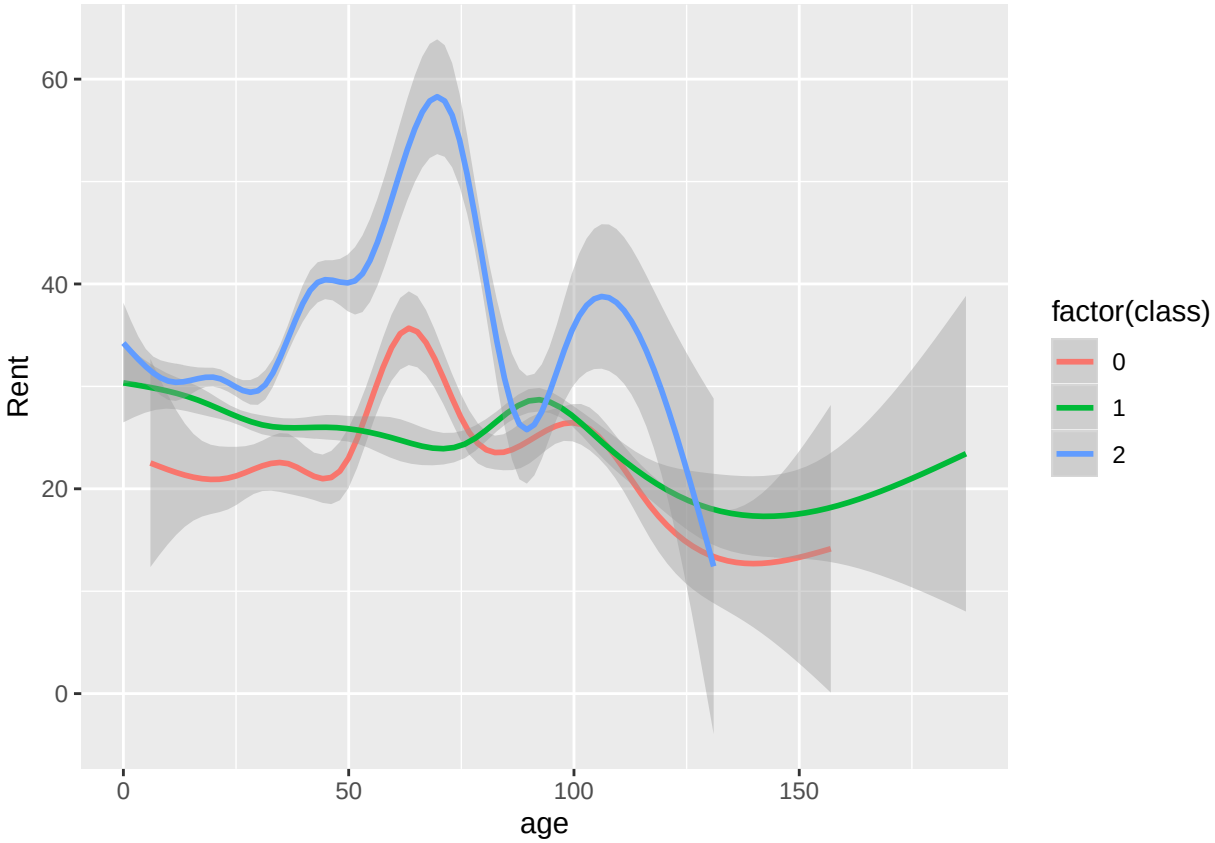
and Class, we construct a scatter plot with age as independent variable, and Rent as dependent variable. But we highlight buildings of different class with different color and attempt to find the impacts of class on rents when we fix the age constant.

```
ggplot(data = greenbuilding) +
  geom_point(mapping = aes(x = age, y = Rent, color = factor(class)))
```



The above scatter plot shows that most buildings labelled as class A are typically newer, while buildings labelled as class B and class C are more widespread in terms of their age distribution. We generate three fitted curves to see how the rents of buildings of different classes change according to their ages.

```
ggplot(data = greenbuilding) +
  geom_smooth(mapping = aes(x = age, y = Rent, color = factor(class)))
```



The above graph reflects the relationship between rent and age for three classes of buildings. It reflects that, as expected, class a buildings are generally more expensive than class b, which is more expensive than class c. We observe some anomalous features of this graph, such as the fact that class c buildings (which we identified as all those that are not designated as class a or class b), command higher rents than class b buildings at various points in their life cycle. One possible explanation is that our assumption about undesignated properties being class c designation may not be reasonable. It also reflects that rents for class a buildings increase from age 25 to about 75, which seems counterintuitive. But the anomalies may be due to the existence a few outlines.

```
xtabs(~class + green_rating, data = greenbuilding)
```

```
##      green_rating
## class    0    1
##     0 1015    7
##     1 3391  131
##     2 2589  546
```

```
p_classa_green = 546/(131+7+546)
p_classa_nongreen = 2589/(3391+2589+1015)
P_classb_green = 131/(131+7+546)
P_classb_nongreen = 3391/(3391+2589+1015)
P_classc_green = 7/(131+7+546)
P_classc_nongreen = 1015/(3391+2589+1015)
p_classa_green
```

```
## [1] 0.7982456
```

```
p_classa_nongreen
```

```
## [1] 0.3701215
```

```
P_classb_green
```

```
## [1] 0.1915205
```

```
P_classb_nongreen
```

```
## [1] 0.4847748
```

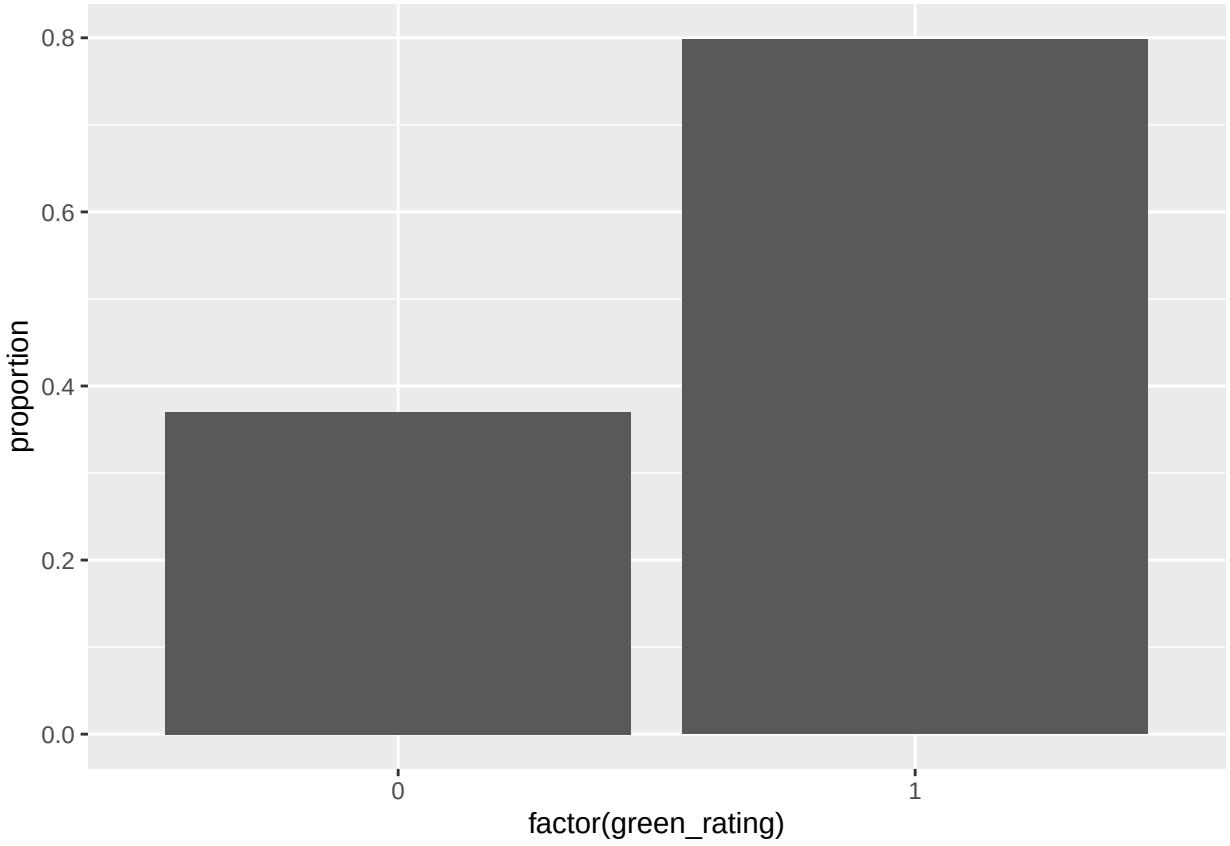
```
P_classc_green
```

```
## [1] 0.01023392
```

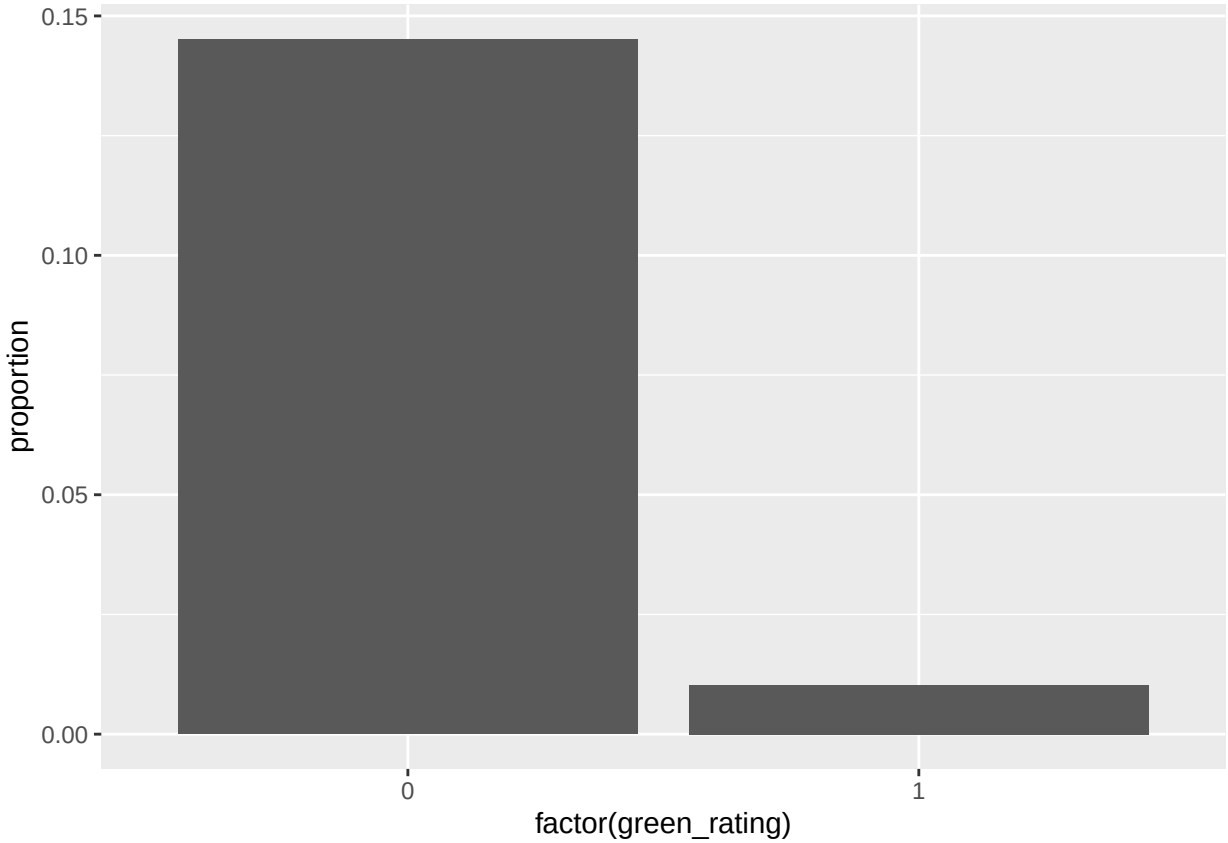
```
P_classc_nongreen
```

```
## [1] 0.1451036
```

```
d1 = greenbuilding %>%
   group_by(green_rating) %>%
   summarize(proportion = sum(class==2)/n())
d2 = greenbuilding %>%
   group_by(green_rating) %>%
   summarize(proportion = sum(class==0)/n())
ggplot(data = d1) +
  geom_bar(mapping = aes(x=factor(green_rating), y=proportion),
           position="dodge", stat='identity')
```




```
ggplot(data = d2) +
  geom_bar(mapping = aes(x=factor(green_rating), y=proportion),
           position="dodge", stat='identity')
```



The tabulation shows that an overwhelming majority of green buildings are classified as class A. While only 37% of the non-green buildings are of class A, 79.8% of the green buildings are class A buildings. Also, only 1% of green buildings are considered least desirable. 14.5% of non-green buildings, however, are deemed as least desirable. With the fact that buildings of a higher class tend to charge a higher rent than those of a lower class, we believe the effect of green certification on rents are greatly overestimated. The following regression results also back up our conclusions. When we add class_a as one of the explanatory variables, the average effect green certifications have on rents drop from 0.545 to -1.135. This not only suggests the importance of class in determining the rents, it also demonstrates the strong correlation between green rating and class, which is ignored by the previous data analyst. After adding this new variable, it becomes apparent that green buildings not only doesn't help increase the rents, it actually contributes to a decrease in the overall rents.

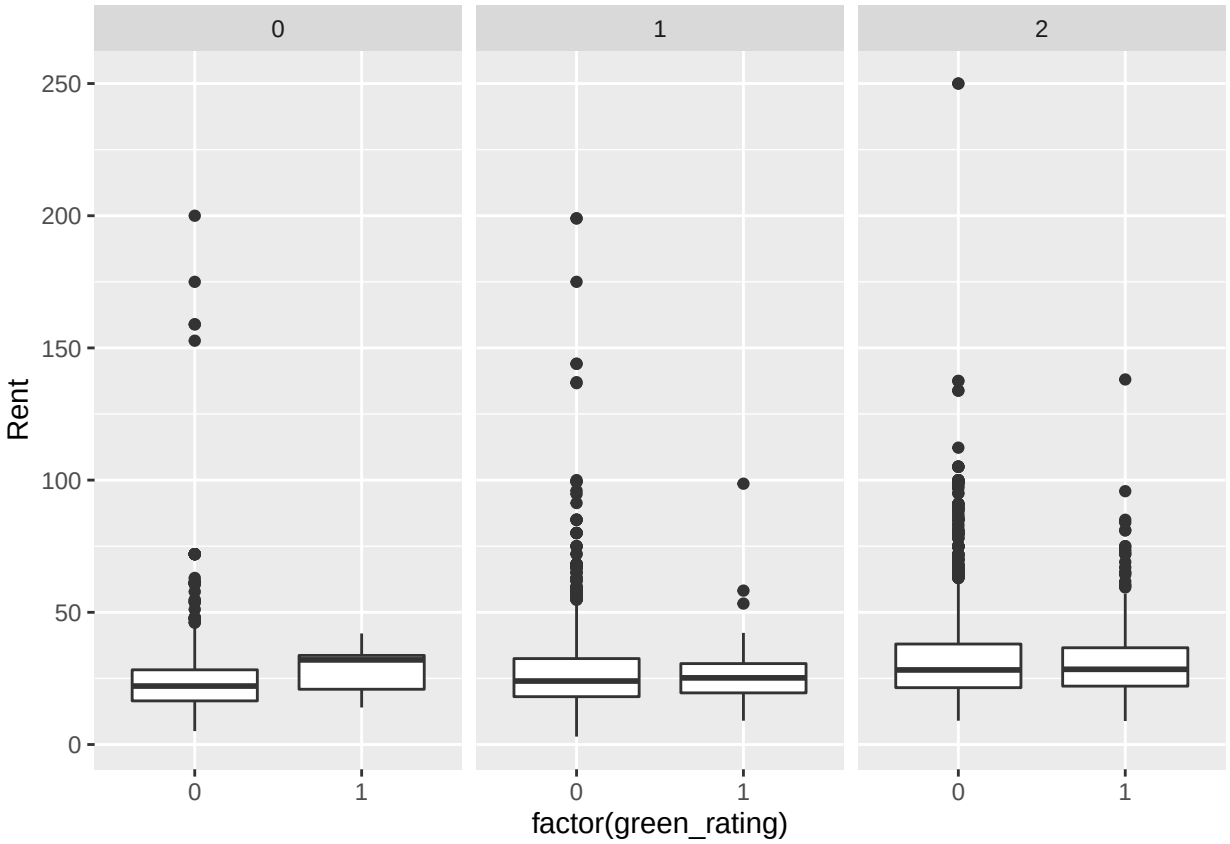
```
lm2 = lm(Rent ~ age + green_rating, data=greenbuildings)
coef(lm2)
```

```
##  (Intercept)          age green_rating
##  30.59205529  -0.04700633   0.54485180
```

```
lm3 = lm(Rent ~ age + green_rating + class_a, data=greenbuildings)
coef(lm3)
```

```
##  (Intercept)          age green_rating      class_a
## 25.782505275  0.001161705 -1.134964372  6.700438077
```

```
ggplot(data = greenbuilding) +
  geom_boxplot(mapping=aes(x=factor(green_rating), y=Rent)) +
  facet_wrap(~class)
```



To make an initial assessment of the economic value of a green building designation, we have now compared rents on green buildings vs. other buildings, separated by class (0 is class C, 1 is class B and 3 is class A). The above box plot reflects the mean and first standard deviations of rent received for such buildings (0 is non-green, 1 is green). This analysis suggests that, controlling for class only, rent is not materially different for green and non-green buildings. There is a somewhat higher mean value for green class c buildings, but we suspect that is an aberration resulting from the relatively small data set for such properties.

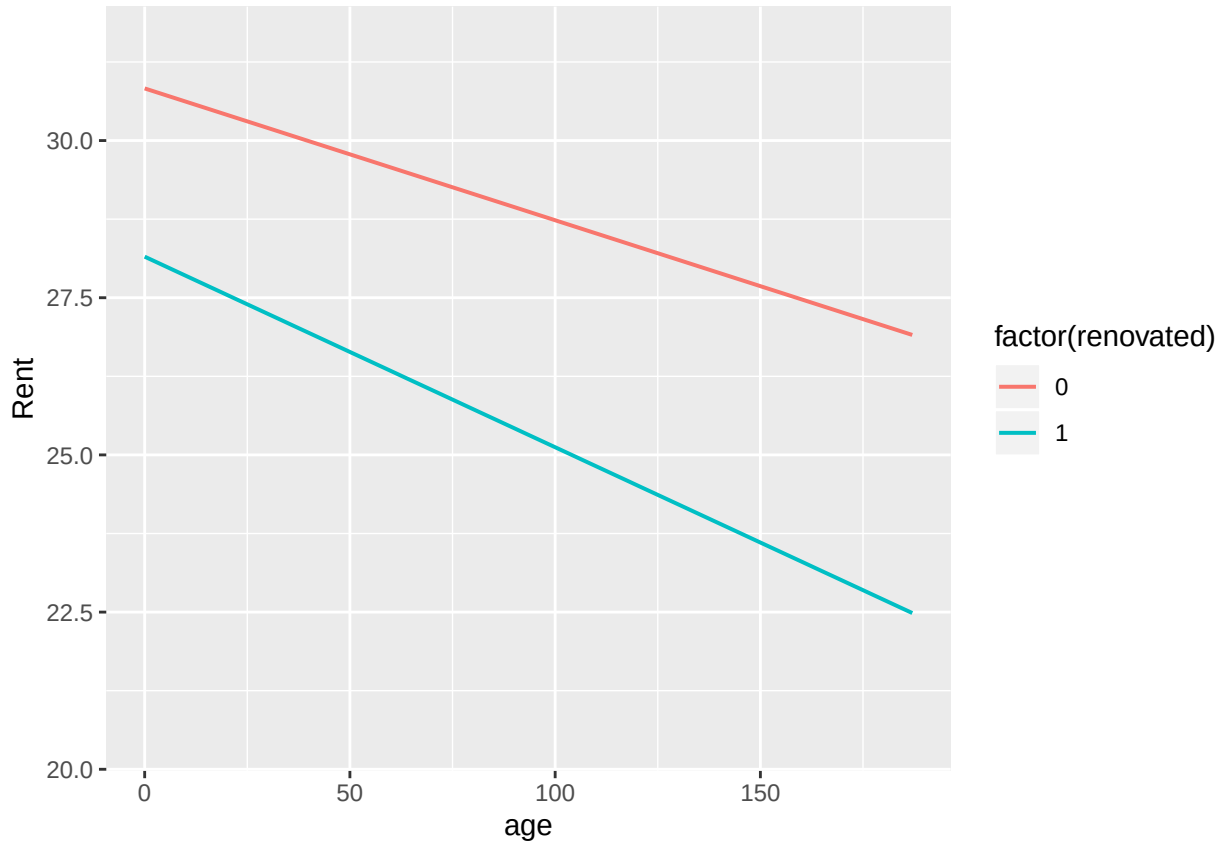
Below we continue to assess separately a variety of potentially confounding factors that may be relevant to our analysis of rents.

Rent & Renovated

```
ggplot(data = greenbuilding) +
  geom_point(mapping = aes(x = age, y = Rent, color = factor(renovated)))
```

Generally, we believe older buildings are more likely to have undergone substantial renovations. The preceding plot backs us up on that. Buildings that have had renovations are typically 30 years old and above. Therefore, to look at the effect renovations have on rents, we generate two separate fitted lines, each of them shows the relationship between age and rents. However, one of the line is generated from buildings that have undergone major renovations, while the other one is generated from buildings that have not.

```
ggplot(data = greenbuilding) +
  geom_lm(mapping = aes(x = age, y = Rent, color = factor(renovated)))
```



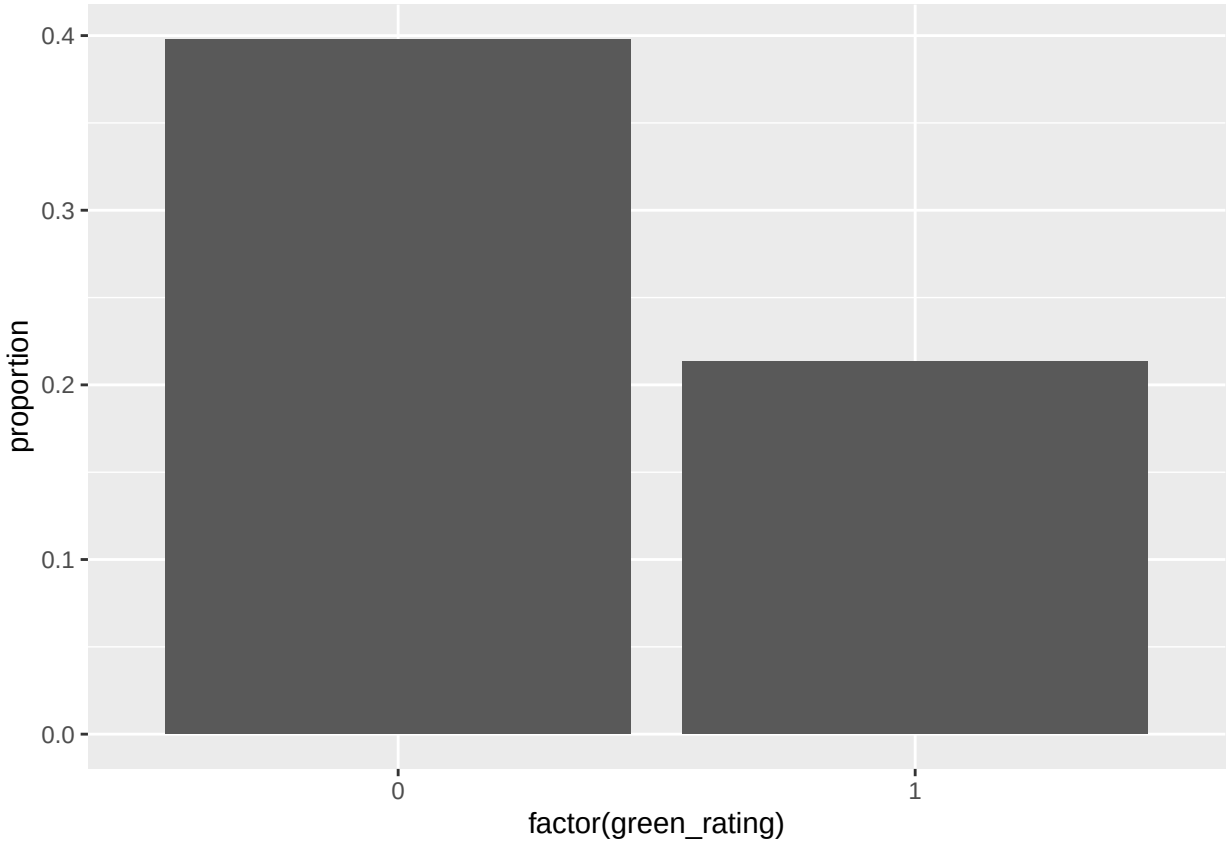
From the fitted lines above, we observe that buildings that haven't undergone substantial renovations tend to have higher rents than those that have.

Next, we investigate the relationship between renovations and green_rating.

```
xtabs(~renovated + green_rating, data = greenbuilding)
```

```
##          green_rating
## renovated    0    1
##         0 4212  538
##         1 2783  146
```

```
d3 = greenbuilding %>%
   group_by(green_rating) %>%
   summarize(proportion = sum(renovated == 1)/n())
ggplot(data = d3) +
  geom_bar(mapping = aes(x=factor(green_rating), y=proportion),
           position="dodge", stat='identity')
```

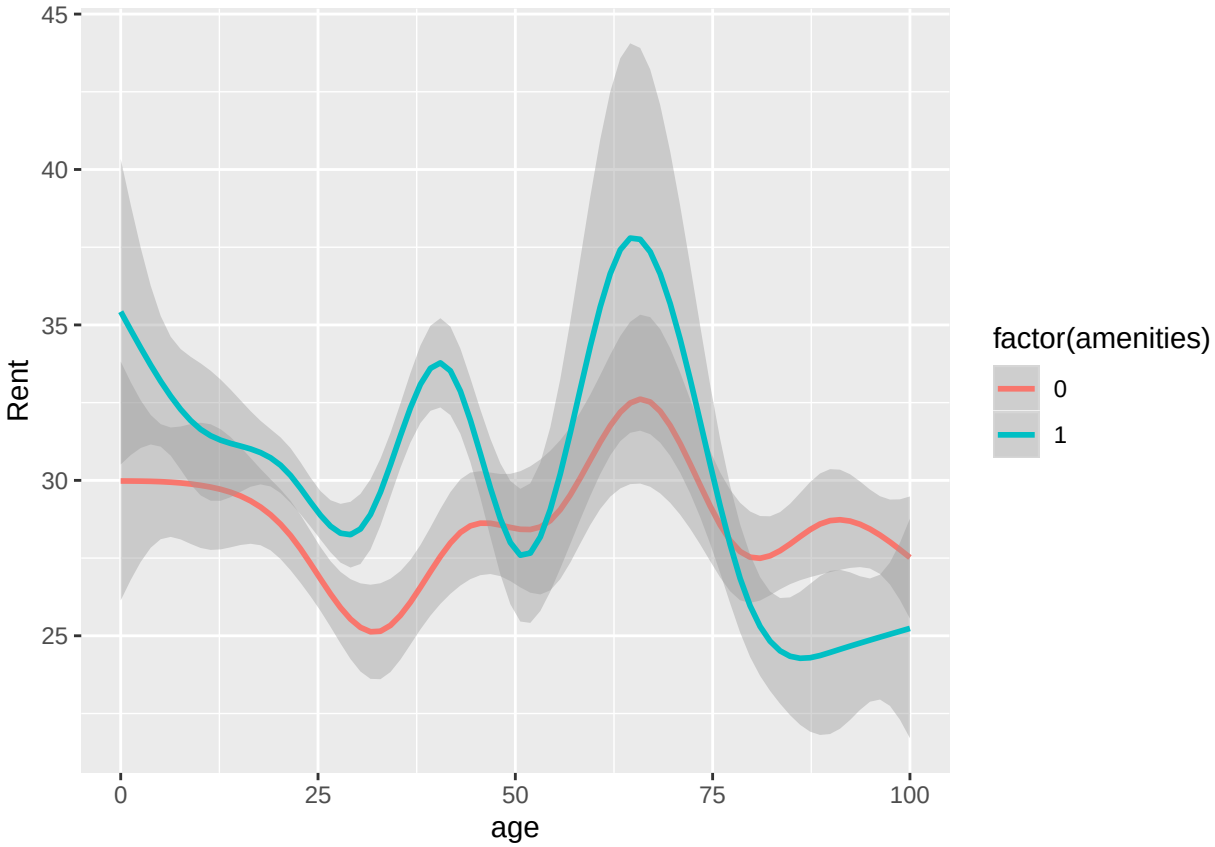


The bar plot above shows that approximately 20% of the green buildings have undergone substantial renovations. Yet nearly 40% of non-green buildings have undergone renovations on a similar scale. Our previous findings show that rents for buildings that have had renovations tend to be lower. Namely, part of the reasons non-green buildings have lower rents is that more non-green buildings have had renovations, which negatively affects the rents. In other words, the effect green certifications have on rents is, again, overestimated in the previous report.

Rent & Amenities

```
ggplot(data = greenbuilding) +
  geom_point(mapping = aes(x = age, y = Rent, color = factor(amenities)))
```


```
ggplot(data = greenbuilding) +
  geom_smooth(mapping = aes(x = age, y = Rent, color = factor(amenities))) +
  xlim(0, 100)
```



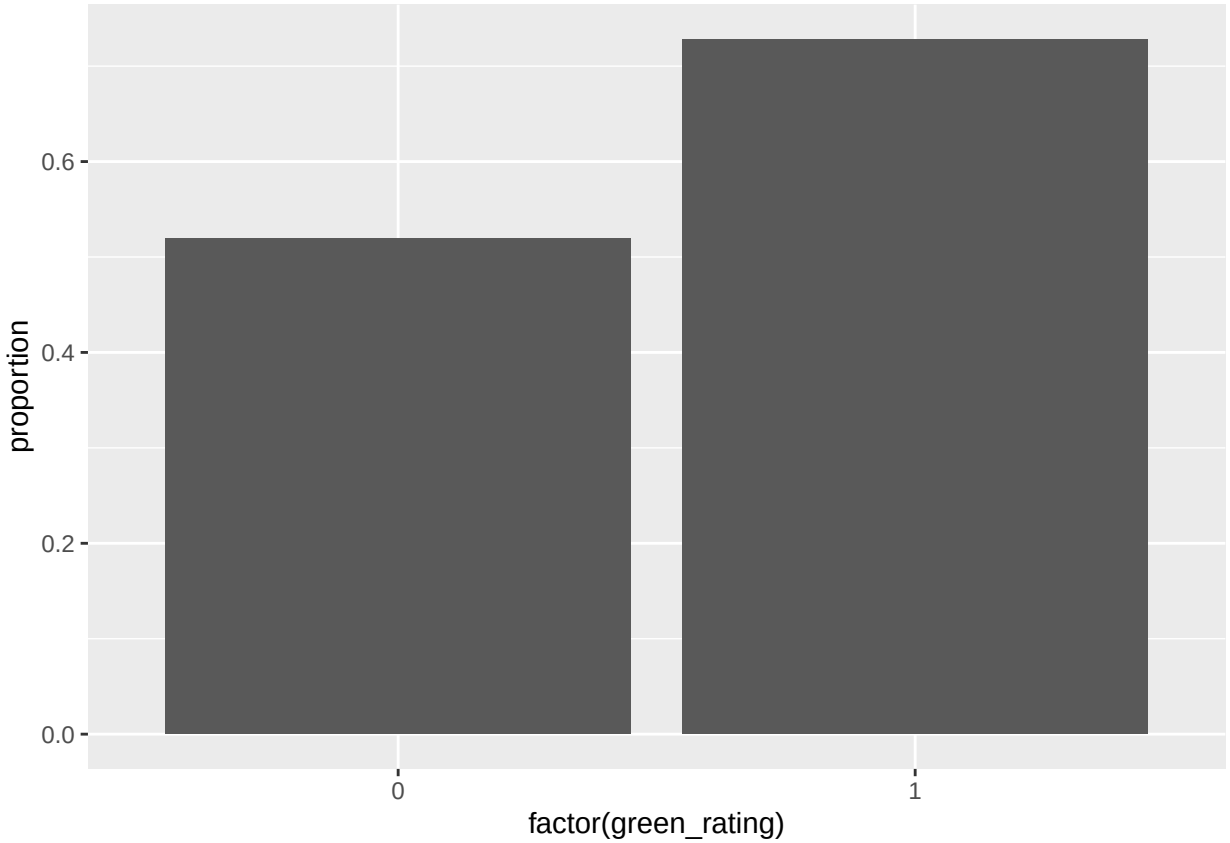
Using similar methodologies, we examine the effect amenities have on rents. The scatter plot shows that newer buildings tend to offer amenities, which we suspect is due to a new trend in architecture design. Again, to examine the effect of amenities, we fix the age variable. Due to the fact that very few buildings over 100 years old have amenities, we restrict the scope of our model to within 100 years. The fitted curve shows that amenities do seem to positively impact the rent, at least when buildings are less than 75 years old.

We further investigate the relationship between green_rating and amenities.

```
xtabs(~amenities + green_rating, data = greenbuilding)
```

```
##           green_rating
## amenities    0    1
##         0 3362  186
##         1 3633  498
```

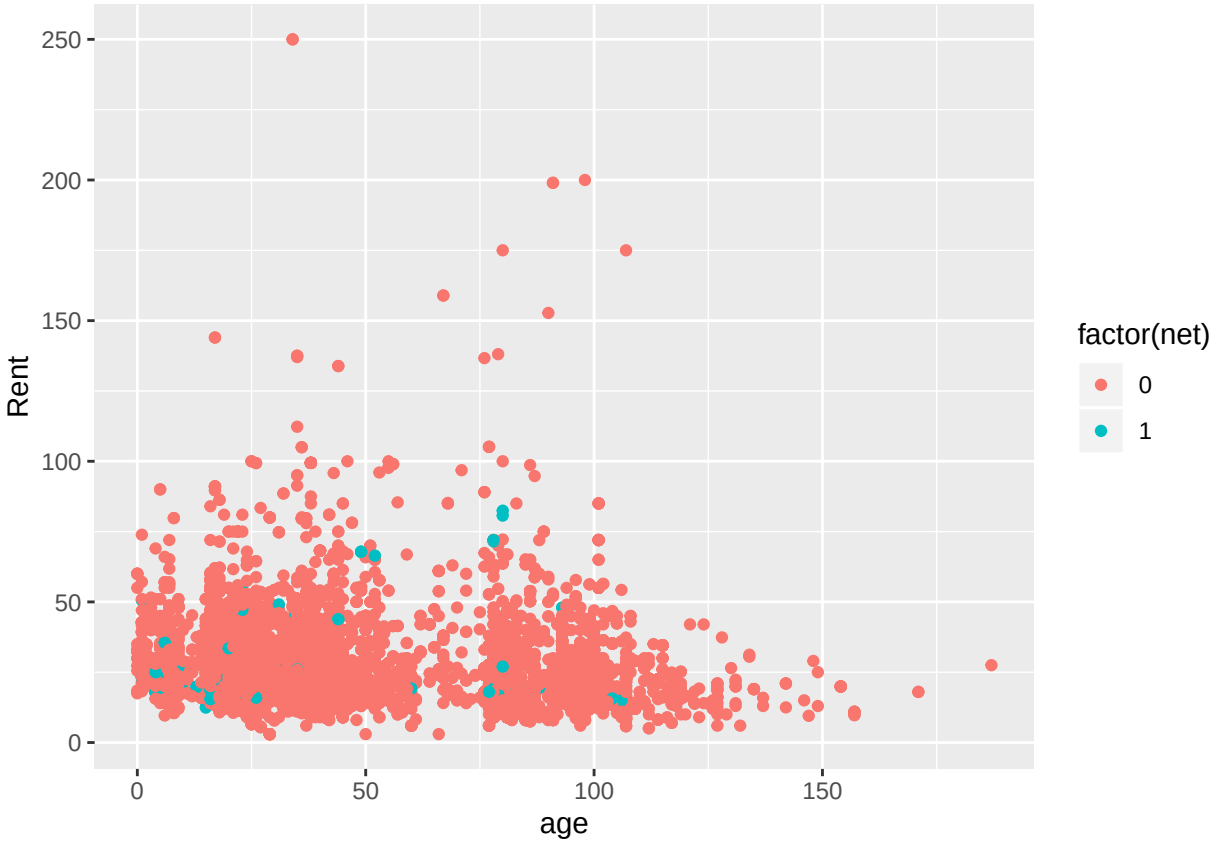
```
d4 = greenbuilding %>%
   group_by(green_rating) %>%
   summarize(proportion = sum(amenities == 1)/n())
ggplot(data = d4) +
  geom_bar(mapping = aes(x=factor(green_rating), y=proportion),
           position="dodge", stat='identity')
```



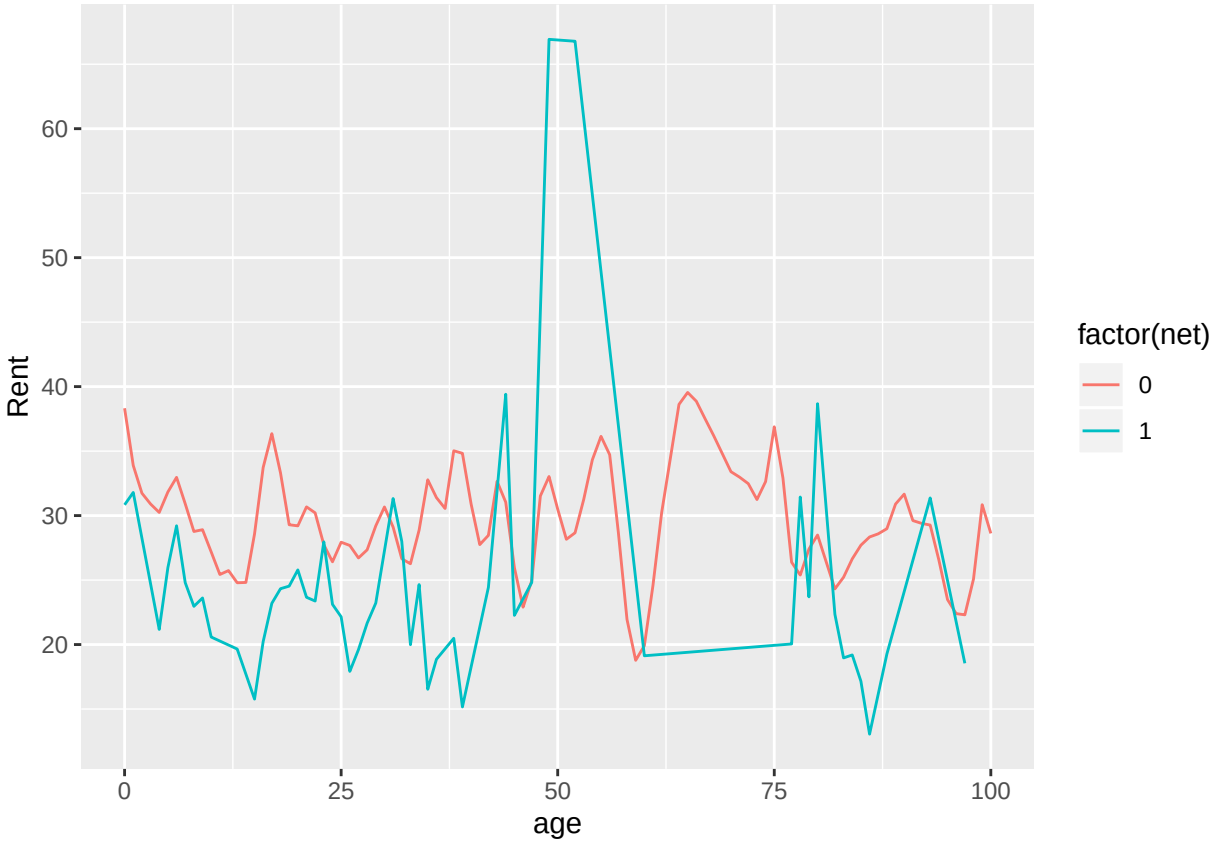
The above bar plot shows that green buildings are more likely to have amenities. Over 70% of green buildings come with amenities, while only around 50% of green buildings have amenities. Combined with the fact that amenities do positively impact rents, we believe part of the effect green certification has on rents actually come from amenities. Green buildings means a higher possibility of having amenities, which increase rents. Therefore, we can conclude that the effect of green certification is, once again, overestimated in the previous report.

Rent & Net

```
ggplot(data = greenbuilding) +
  geom_point(mapping = aes(x = age, y = Rent, color = factor(net)))
```




```
ggplot(data = greenbuilding) +
  geom_spline(mapping = aes(x = age, y = Rent, color = factor(net))) +
  xlim(0, 100)
```



The two diagrams above shows that the rents for buildings with net-rental contracts are typically lower than those without such contracts. One exception occurs around the age of 50, where a few unusually expensive net-rental properties cluster. But overall, net-rental contracts do seem to be associated with lower rents.

```
xtabs(~net + green_rating, data = greenbuilding)
```

```
##    green_rating
## net    0    1
##   0 6761  645
##   1  234   39
```

```
percent_net_green = 39/(39+645)
percent_net_nongreen = 234/(234+6761)
percent_net_green
```

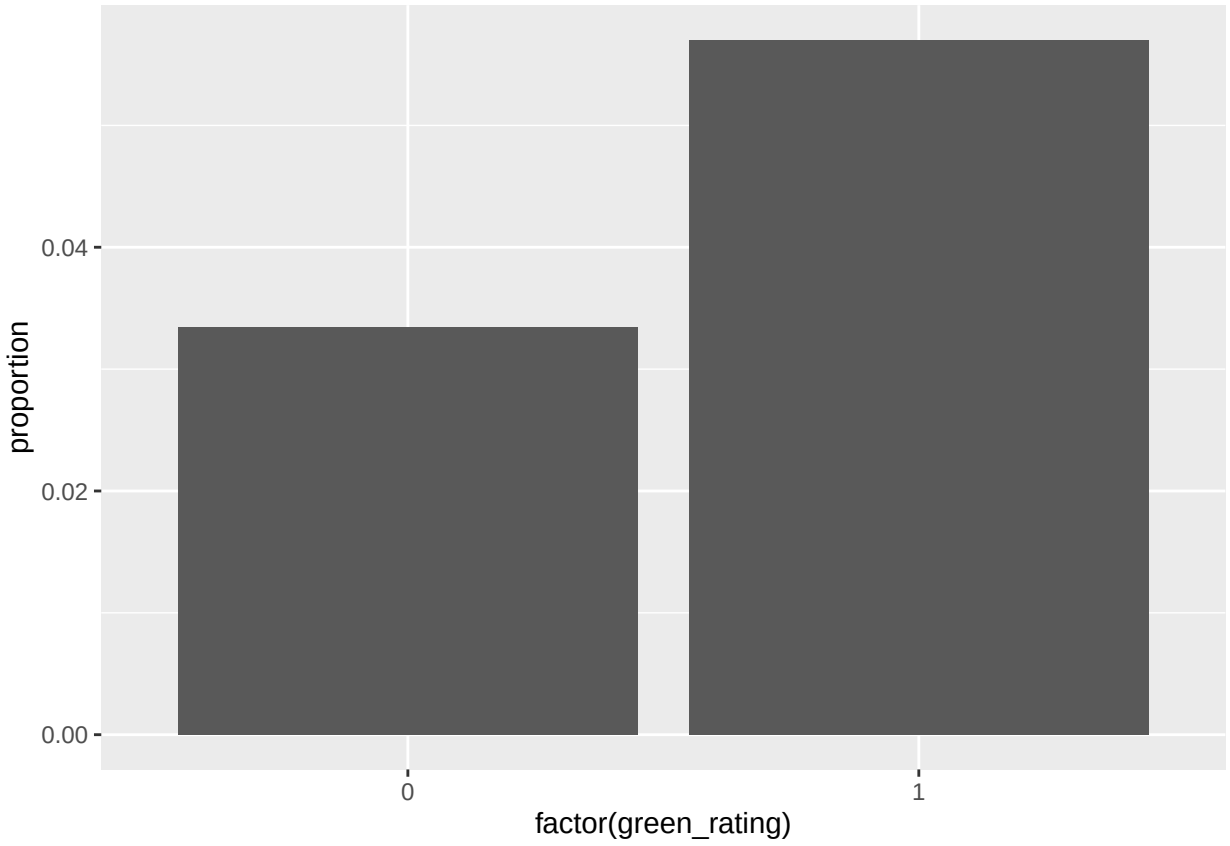
```
## [1] 0.05701754
```

```
percent_net_nongreen
```

```
## [1] 0.03345247
```

```
d5 = greenbuilding %>%
   group_by(green_rating) %>%
   summarize(proportion = sum(net)/n())
```

```
ggplot(data = d5) +
  geom_bar(mapping = aes(x=factor(green_rating), y=proportion),
           position="dodge", stat='identity')
```



We then go on to examine whether the proportions of buildings with net-rental contracts differ between green and non-green buildings. The result shows that 5% of the green buildings employ net-rental contracts, but only 3% of non-green buildings use such contracts. Consequently, we do believe the effect of green certification on rents may be slightly underestimated due to the net variable. But still, due to the relatively minor difference between the two proportions, we still believe the effect of green certifications to be overall overestimated in the previous report.

Final Regression Model

```
lm2 = lm(Rent ~ age + green_rating + net + amenities + renovated + class_a + class_b, data=greenbuildin
coefficients(summary(lm2))
```

```
##                 Estimate  Std. Error   t value      Pr(>|t|)
## (Intercept)  23.22215203 0.638312495 36.380538 4.895797e-268
## age           0.03501298 0.006816361  5.136608  2.864403e-07
## green_rating -1.22306610 0.606462915 -2.016720  4.375853e-02
## net          -6.14394490 0.902926695 -6.804478  1.088347e-11
## amenities    -0.81201182 0.364439034 -2.228114  2.590099e-02
## renovated    -3.65141741 0.386525144 -9.446778  4.514461e-21
```

```
## class_a      10.36035877 0.622022223 16.655930  3.024919e-61
## class_b       3.33371950 0.518605273  6.428241  1.365969e-10
```

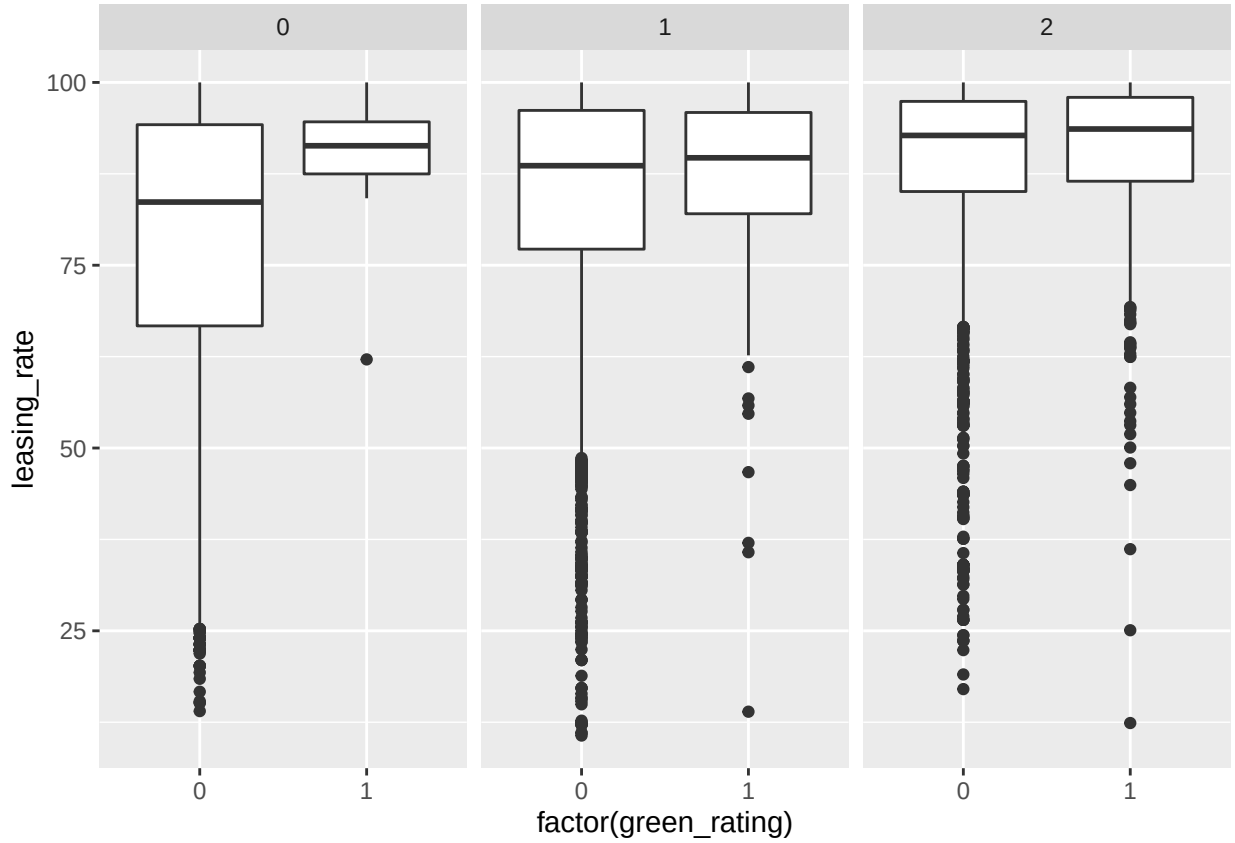
The results from the regression model, which incorporate all variables correlated with rents from the data set, show that green buildings do have a negative impact on rents. This is consistent with our previous analysis, which shows the positive impacts green certifications have on rents are being severely overestimated. From the regression results, we can tell that the rents of green buildings, on average, are $1.223 (per square foot) less than non-green buildings. Therefore, the analyst's story is problematic, and from the perspective of gaining more rents, the green buildings are not advantageous to build.

Occupancy

Another way of looking at profitability of the investment is to examine the occupancy rates. If the occupancy rates of green buildings prove to be higher, some of the negative effects green certifications have on rents may be negated.

As we suspect the occupancy rate to be associated with the buildings' classes, we generate the following box plot representing the relationship between occupancy rates and green certifications, independent of the buildings' classes.

```
ggplot(data = greenbuilding) +
  geom_boxplot(mapping=aes(x=factor(green_rating), y=leasing_rate)) +
  facet_wrap(~class)
```



We compared the impact of green designation on occupancy rates, separated by the building's class. As the box plots above reflect, with the exception of green designated class c buildings (for we suspect is a quirk

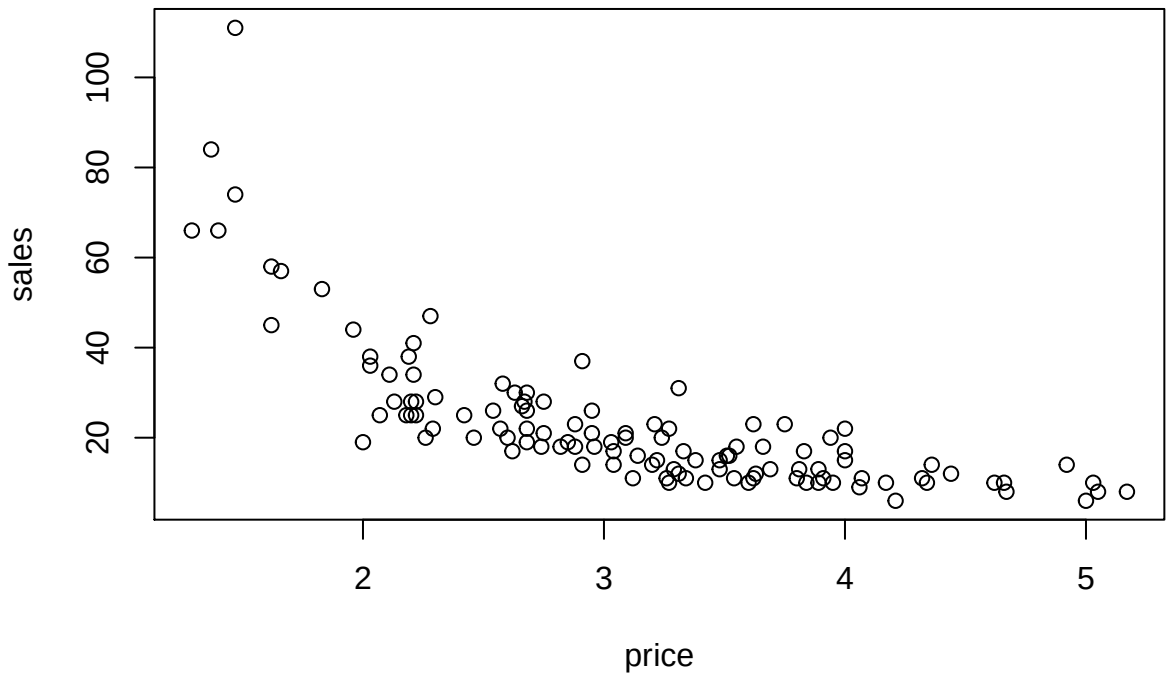
resulting from the fact that there are very few observations), there is little difference between green and non-green designations in terms of occupancy rates.

Conclusion

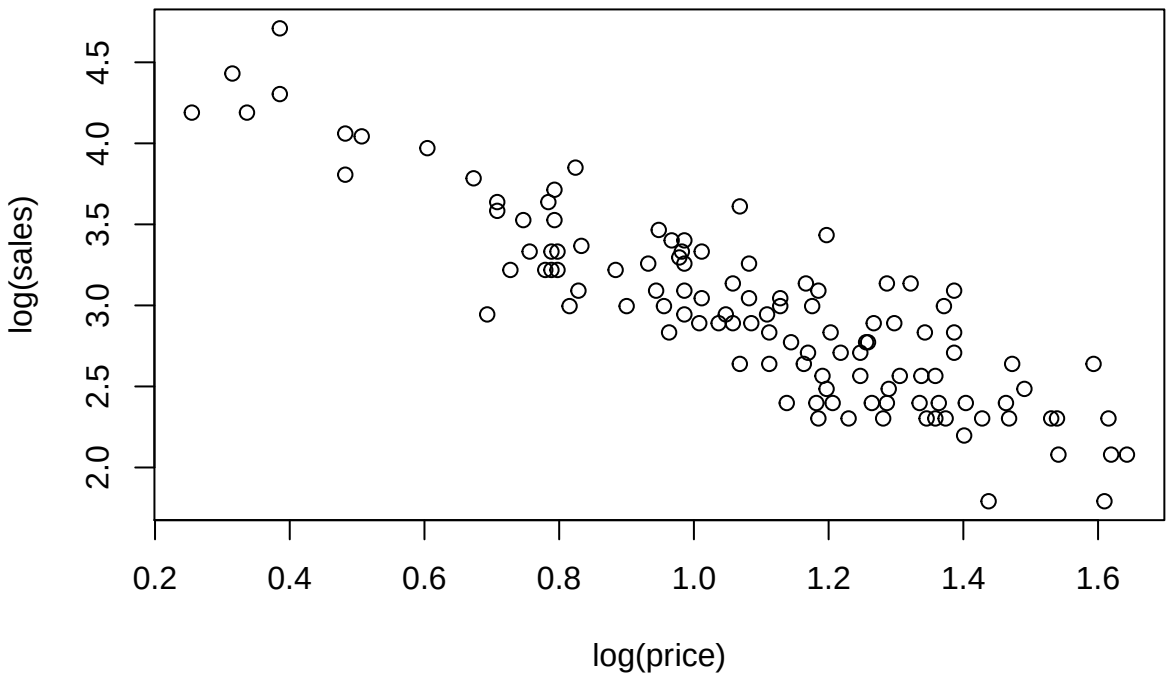
Our assessment of the data does not support the conclusion that investing the additional amounts required to construct a building that will achieve a green designation will be rewarded financially with higher rents or more favorable occupancy, the two most significant drivers of economic return. Accordingly, unless required by regulatory considerations, the decision to make such an investment will have to be supported by qualitative considerations such as concern for the environment and the community.

Milk prices

```
milk <- read.csv("/Users/pengcheng/Desktop/UT Austin/Spring 2020/SDS 323/Exercise 1/milk.csv")
plot(sales ~ price, data=milk)
```




```
plot(log(sales) ~ log(price), data=milk)
```



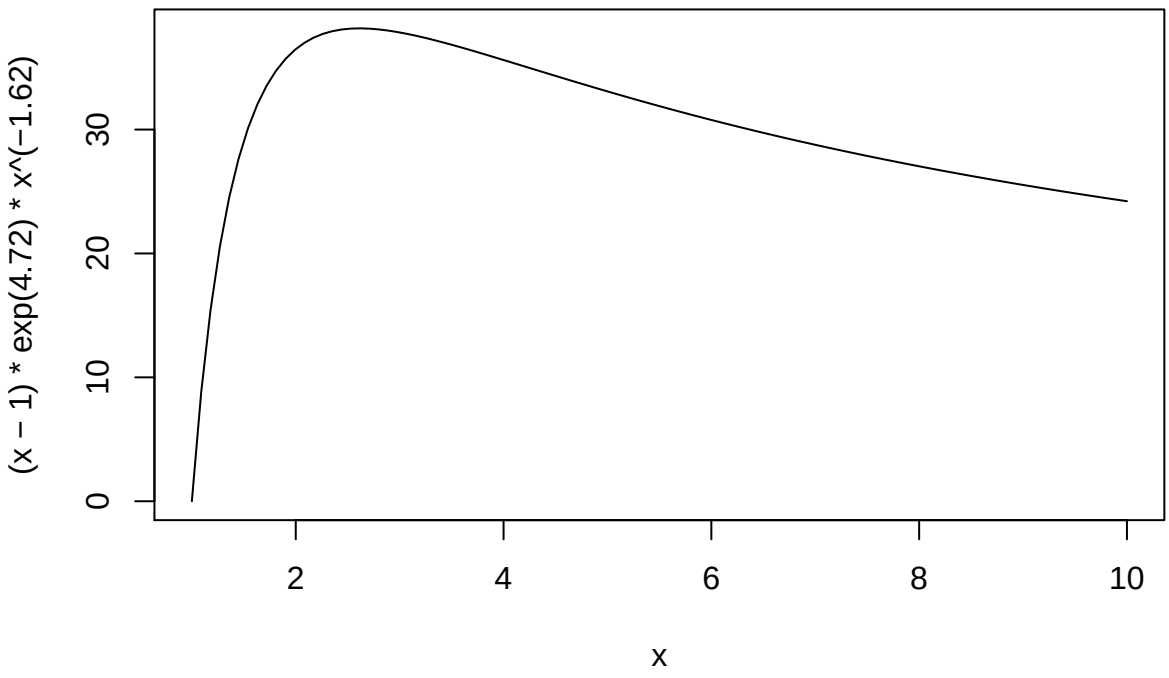

```
lm_price = lm(log(sales) ~ log(price), data=milk)
coefficients(summary(lm_price))
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)  4.720604 0.09171808  51.46863 9.251409e-81
## log(price)  -1.618578 0.08116128 -19.94273 5.692248e-39
```

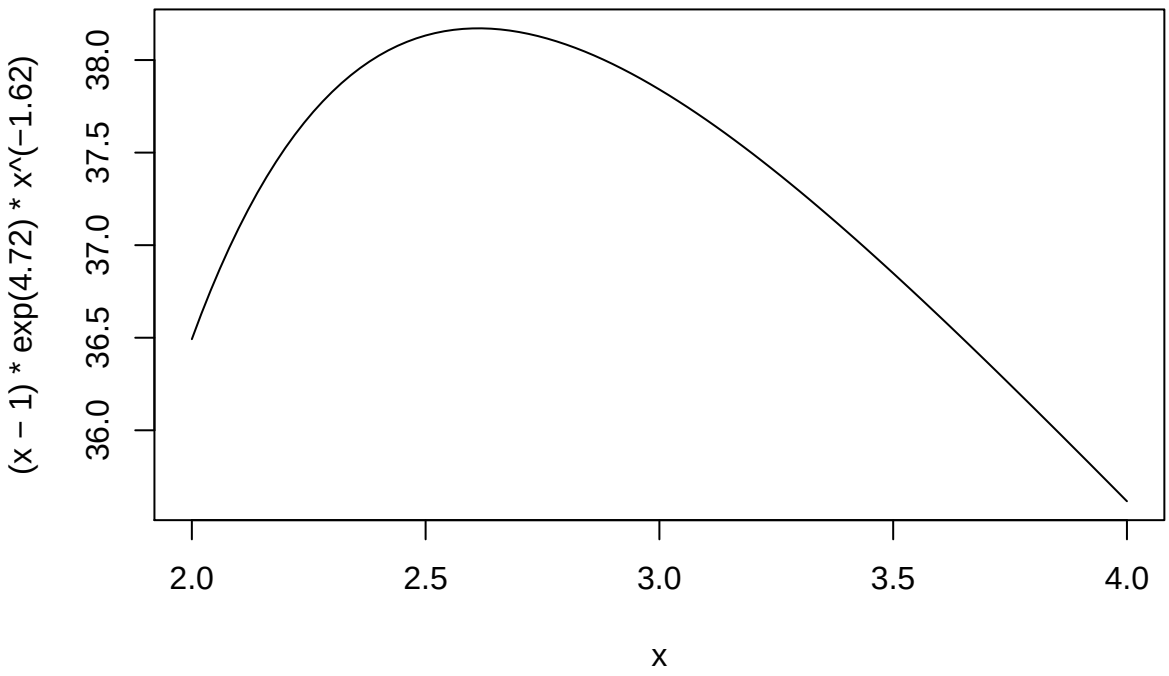
```
x = 2.613
```

The equation calculating the sales is as follows: **sales = exp(4.72) * price ^(-1.62)**

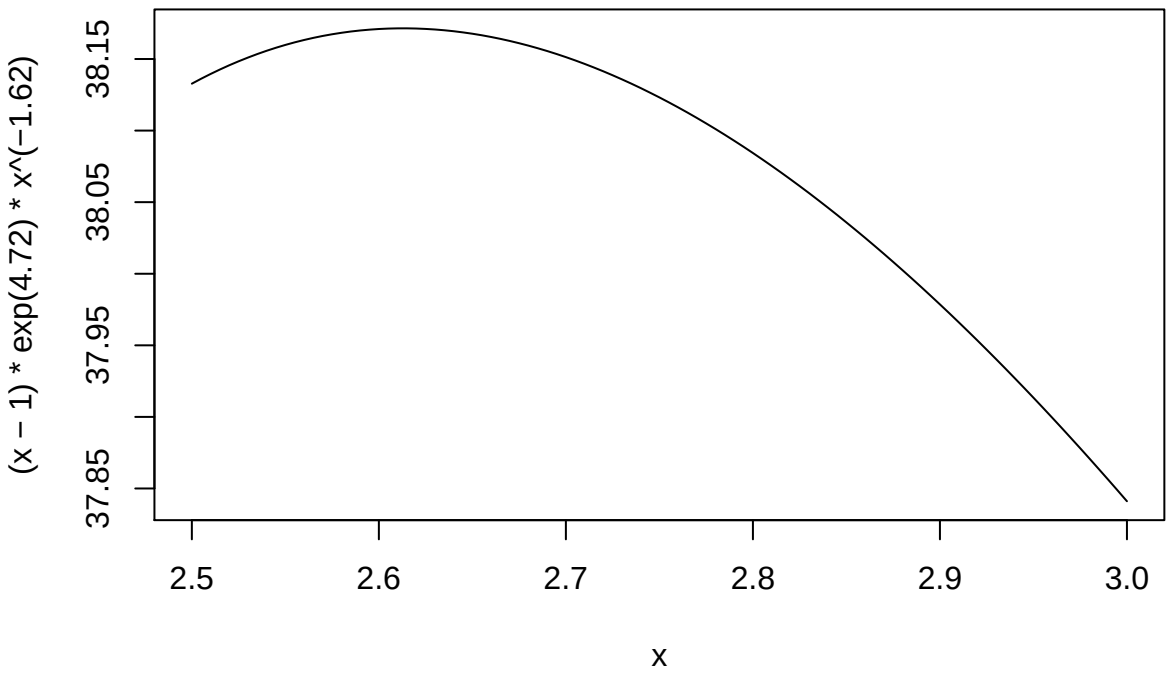
```
curve((x-1)*exp(4.72)*x^(-1.62), from=1, to=10)
```



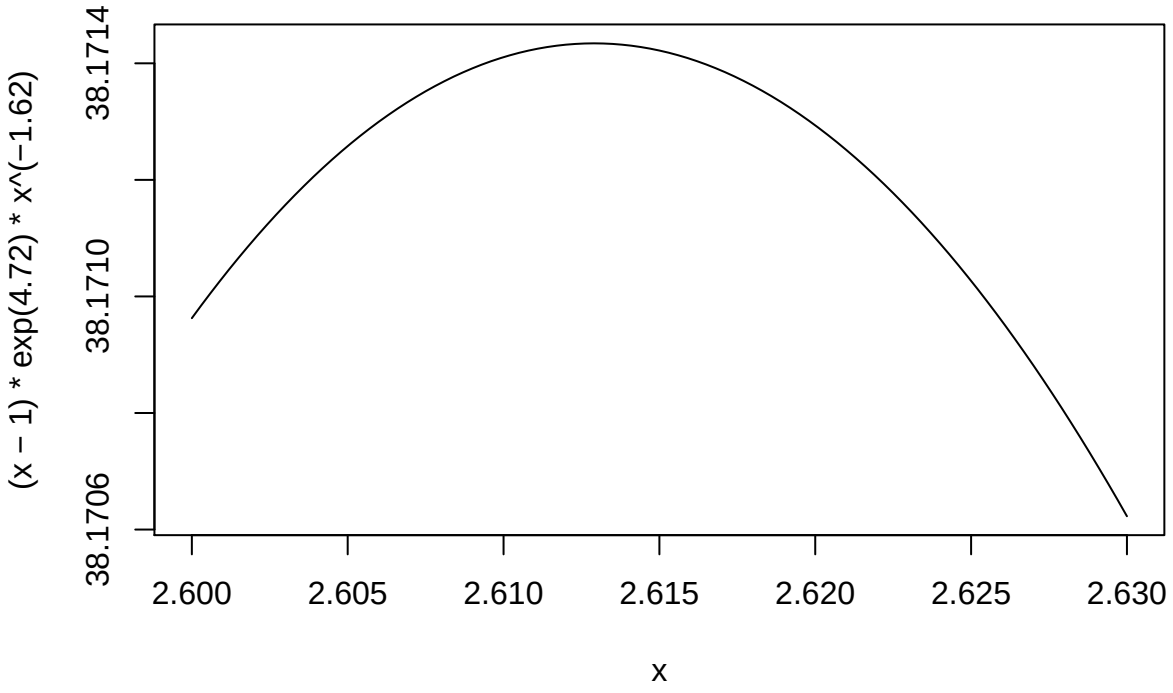

```
curve((x-1)*exp(4.72)*x^(-1.62), from=2, to=4)
```



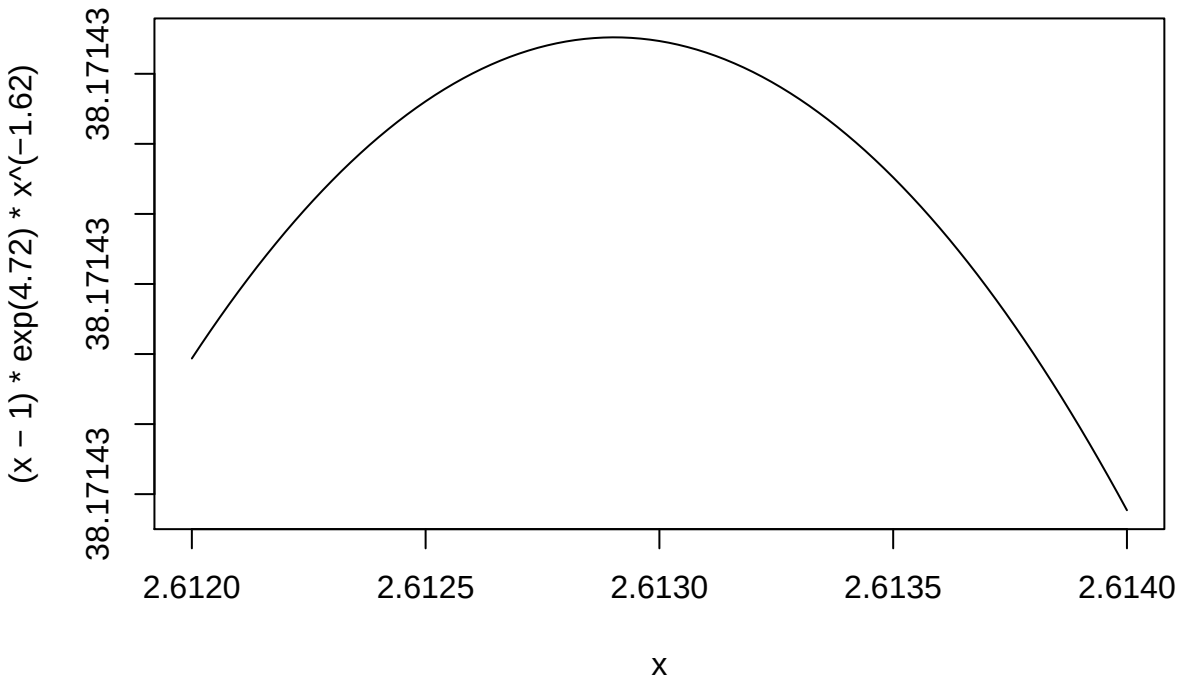

```
curve((x-1)*exp(4.72)*x^(-1.62), from=2.5, to=3)
```




```
curve((x-1)*exp(4.72)*x^(-1.62), from=2.6, to=2.63)
```




```
curve((x-1)*exp(4.72)*x^(-1.62), from=2.612, to=2.614)
```




```
x = 2.613
```

Then, from the graph and calculations above, we can infer that when x, the price of the milk, is $2.613, the revenue can be maximized.