STA 380, Part 2: Exercises

Brandt

8/5/2021

Green Buildings

While the stats guru's analysis provided a good baseline for thinking about the issue, we believe his methodology fell short in several areas. Agree or disagree!? We do not feel that the data provides sufficient evidence, on a monetary basis, to justify the additional investment to construct the building in line with green certification standards.

15 story, east chavex near downtown.

```
##
## -- Column specification --------------------------------------------------------
## cols(
##   .default = col_double()
## )
## i Use `spec()` for the full column specifications.
```

Data Cleaning:

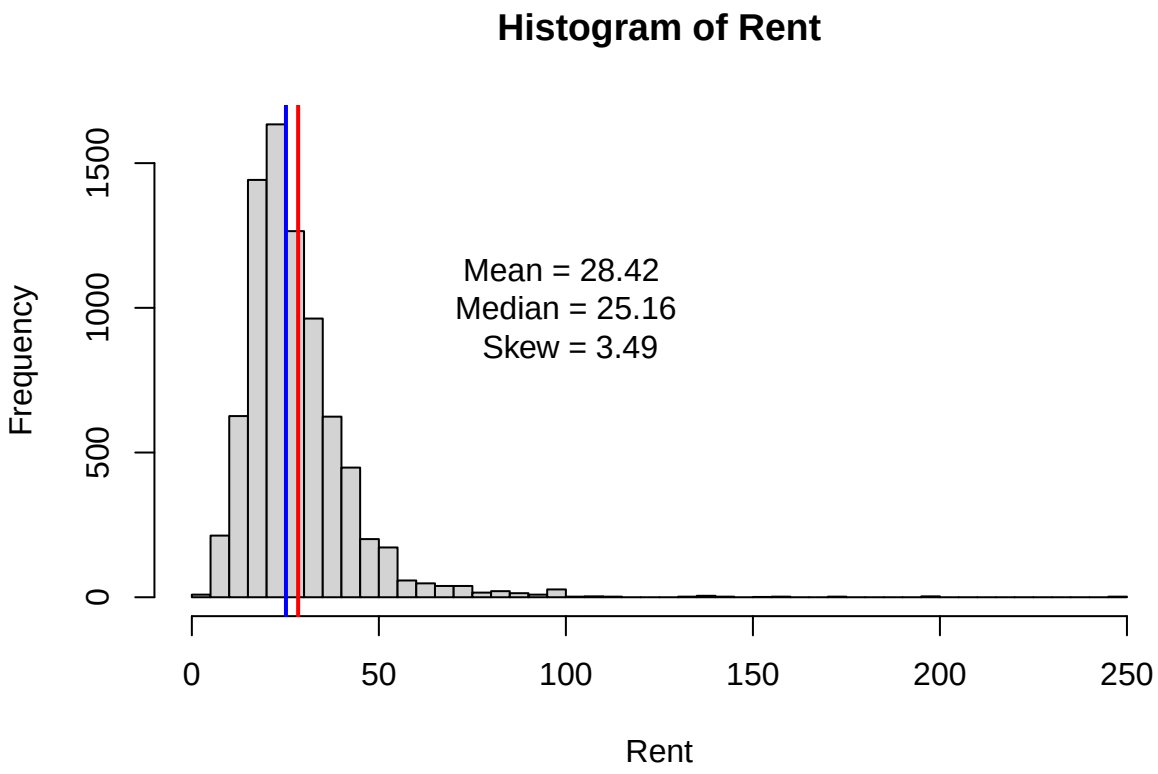
First, we can address the data cleaning aspect of the analysis. Are there any adjustments that should be made? The excel guru certainly felt that a certain amount of scrubbing should take place. He decided that because some of the buildings had low occupancy rates they should be excluded from the analysis due to their "weirdness".

Exploration

Note: For the purposes of presentation clarity, we have converted all categorical, binary variables from "1", "0" to "Yes","No".

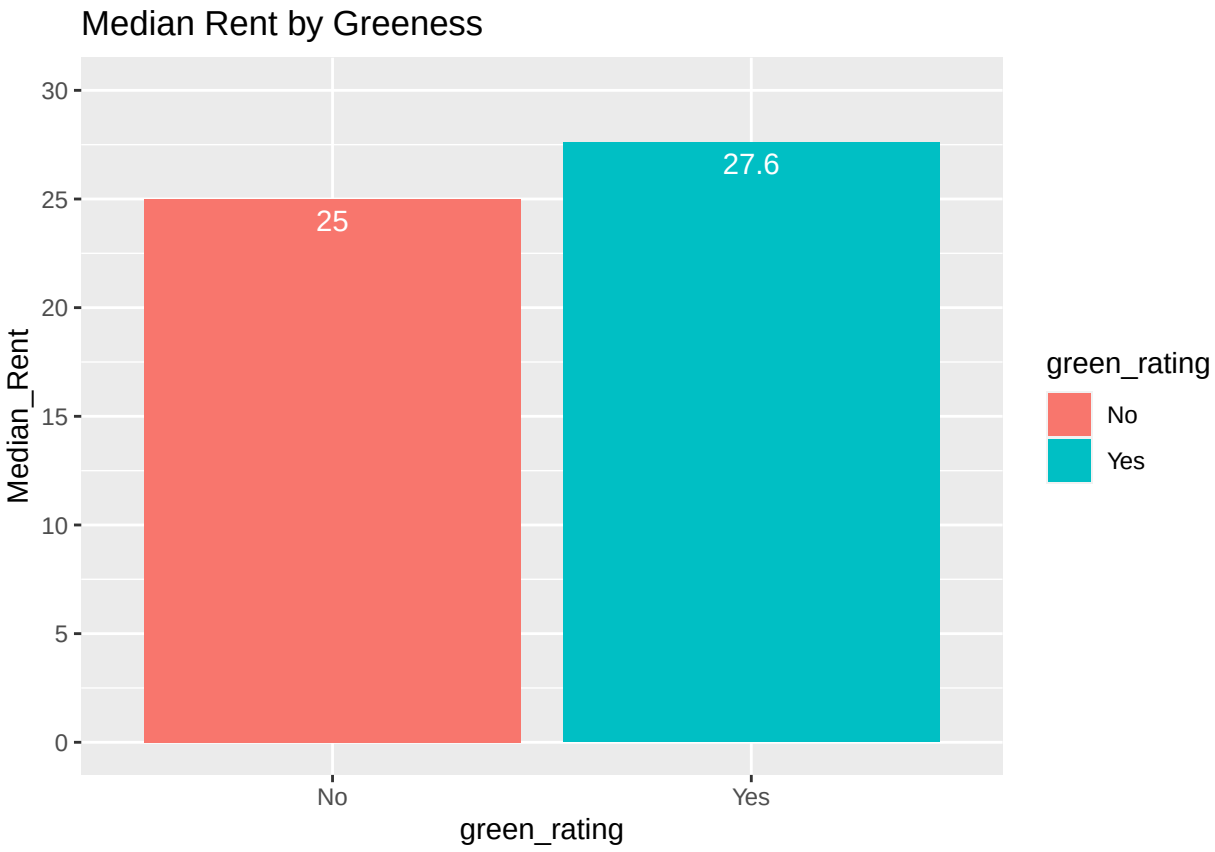
Mean vs Median: Who's in Charge?

The guru decided to use the median, over the mean in his analysis and we feel this approach is justified. Support for this position can be seen in the below histogram which depicts the distribution of rents. The rents are clearly right skewed, with a few points that lie far into the right tail. The vertical, blue line indicates the median value, and the vertical red line indicates the mean. We maintain that the appropriate statistic for our purposes is the median, unless there is some justification or reasoning that would indicate this building is out of the ordinary. From here on, we will confine the majority of our analysis to looking at just the median.



Green vs Non-Green

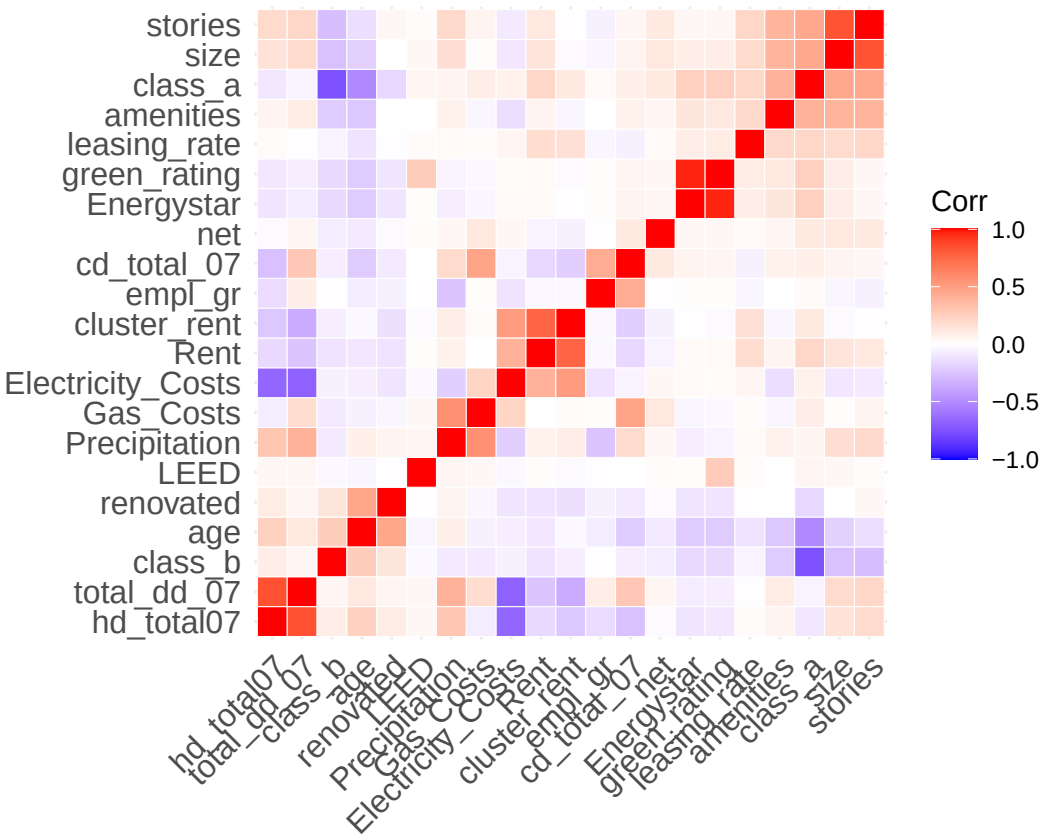
Now, as a starting point we can compare the median rent for green buildings vs non-green buildings.



The median rent is certainly larger for green buildings vs non green buildings: $27.60 per square foot for green buildings vs $25.00 for non-green buildings. But this doesn't really tell the full story. It's too much a leap of faith to claim that this rental difference is due solely to the building's green rating. We need to dig deeper to understand the data further and perhaps discover thatthe higher median rent differential could be attributed to another variable.

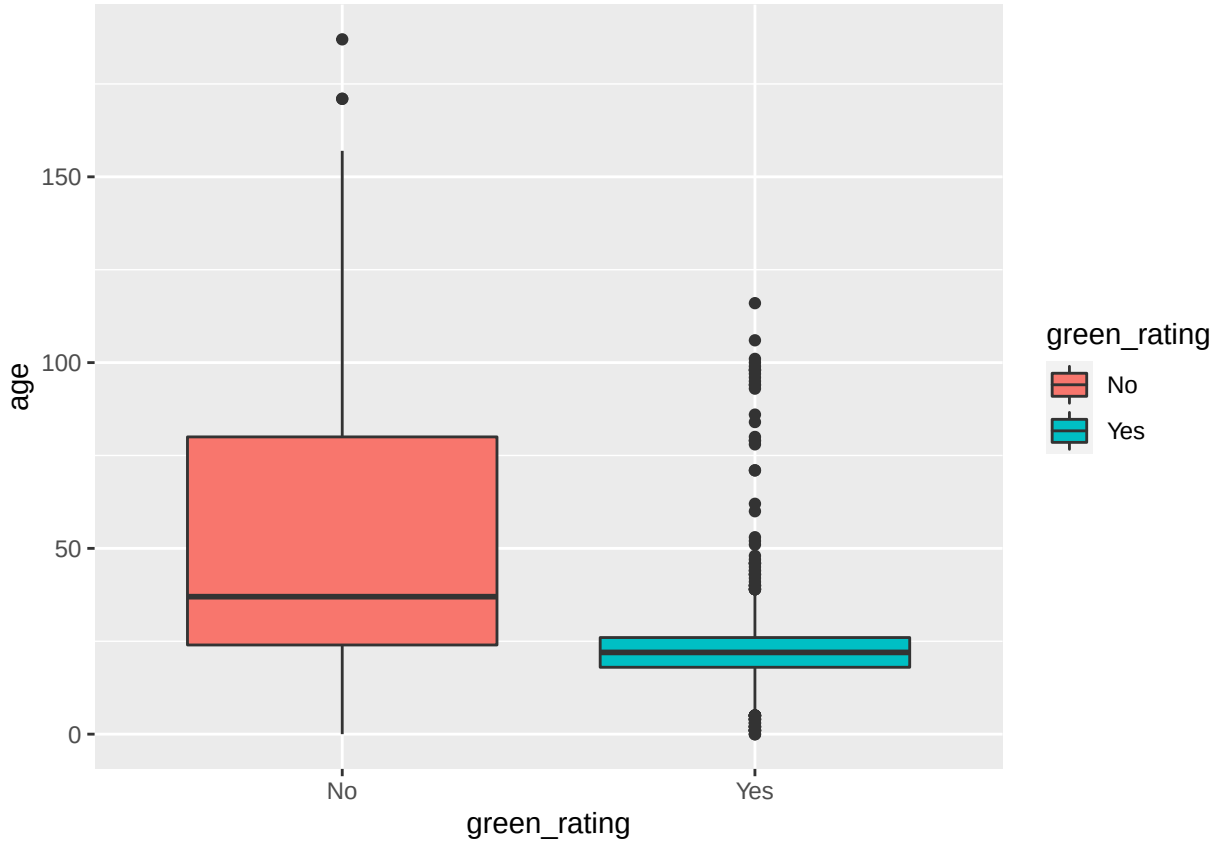
Coundounders

You might hypothesize that green buildings are simply associated with other factors that are really driving the difference in the median rental value. To get a feel for this, we can plot a correlation matrix to understand which variables are correlated with each other.

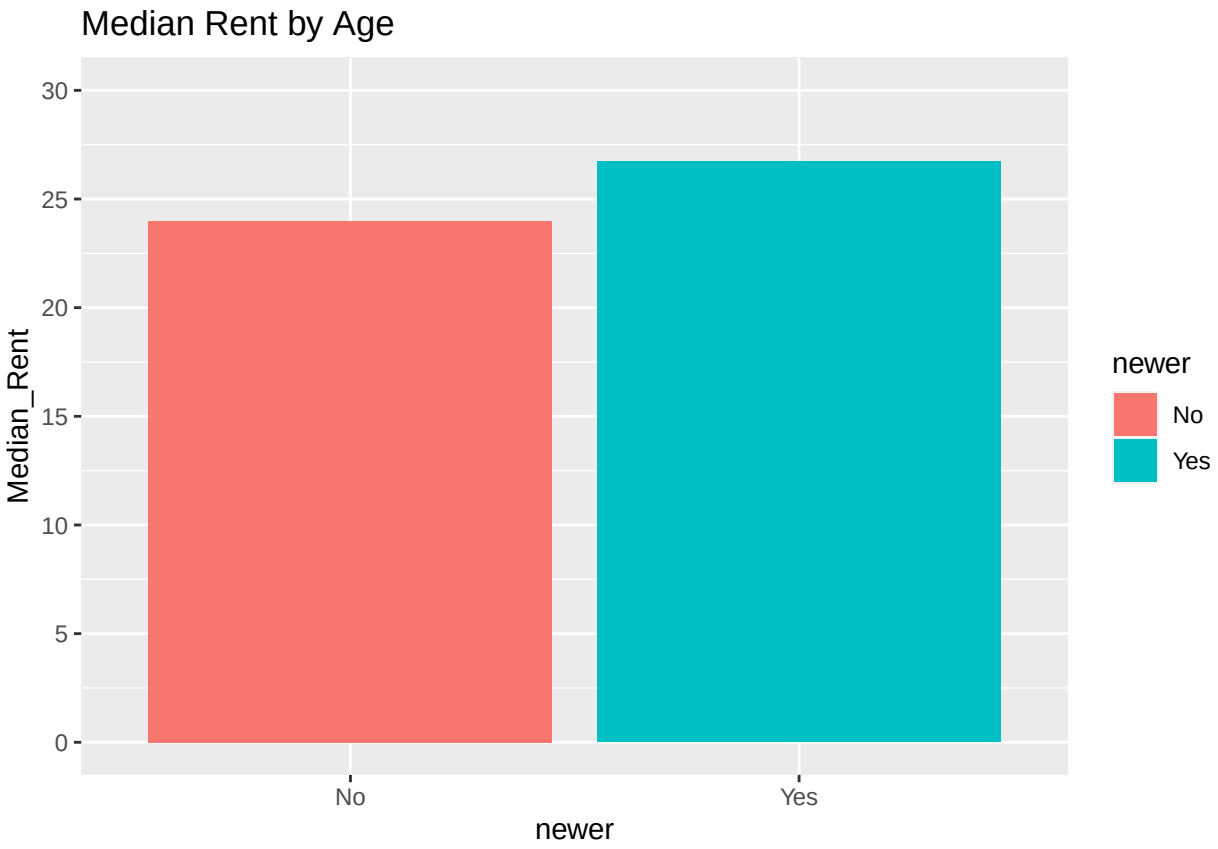


There are lots of interesting relationships displayed here, but we would like to focus on the interplay between Rent, and other variables that may be impacting it. We can see start trying to untangle some of the affects here. We want to examine items that may be correlated with both rent and green ratings.

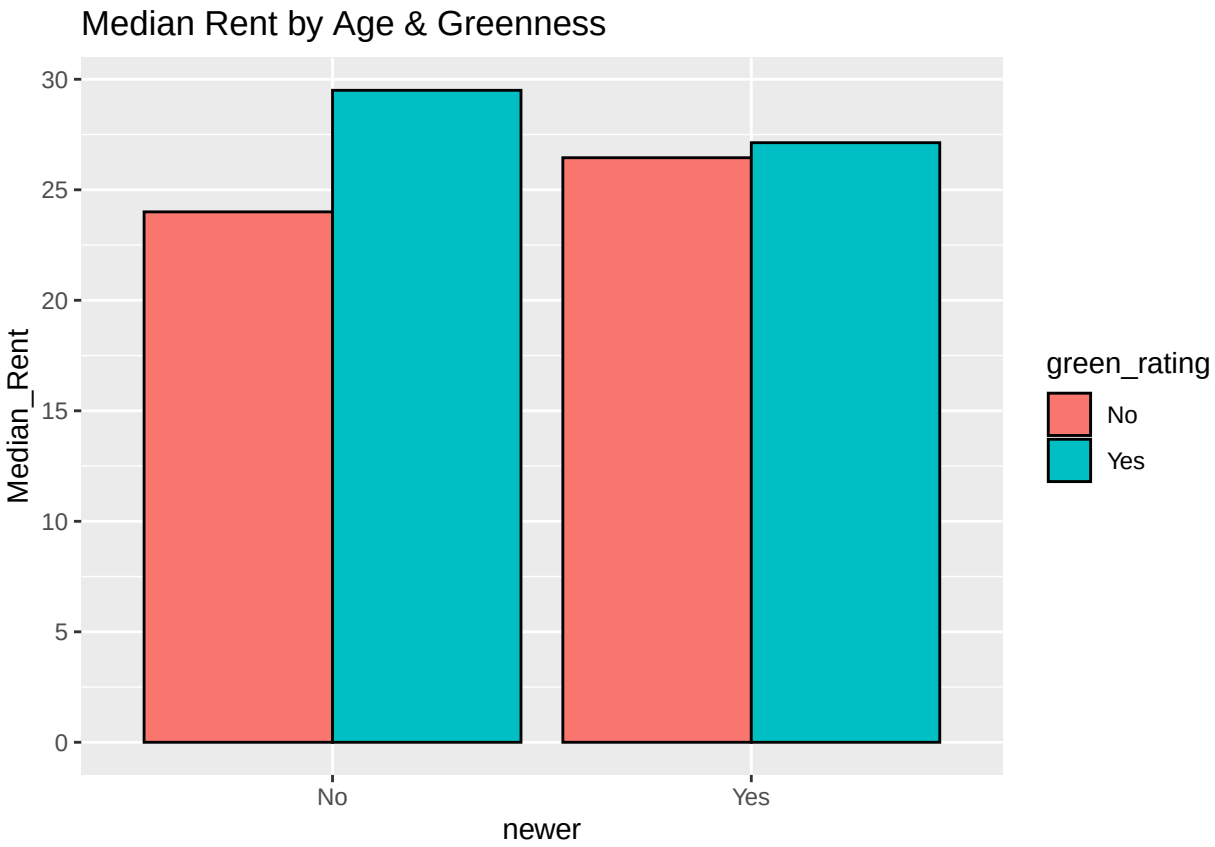
Clearly, greener buildings tend to be much younger.



Yep, newer buildings get higher median rent. But breaking it down more: older green buildings actually have higher rent than younger green buildings. so that's weird. For our purposes, the left table is actually more relevant I would think. Our building will be new!



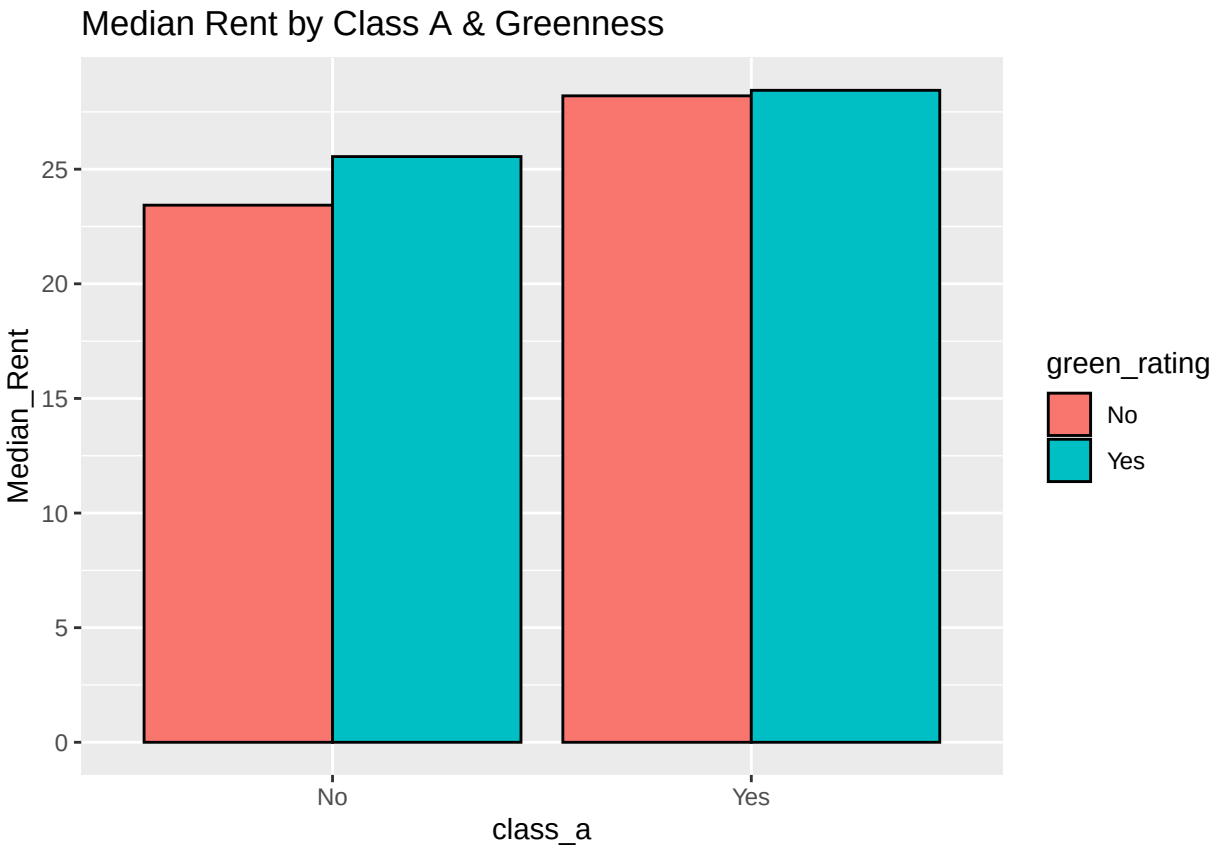

```
## `summarise()` has grouped output by 'green_rating'. You can override using the `.groups` argument.
```



What about class? Perhaps green buildings just happen to be built of 'higher class'

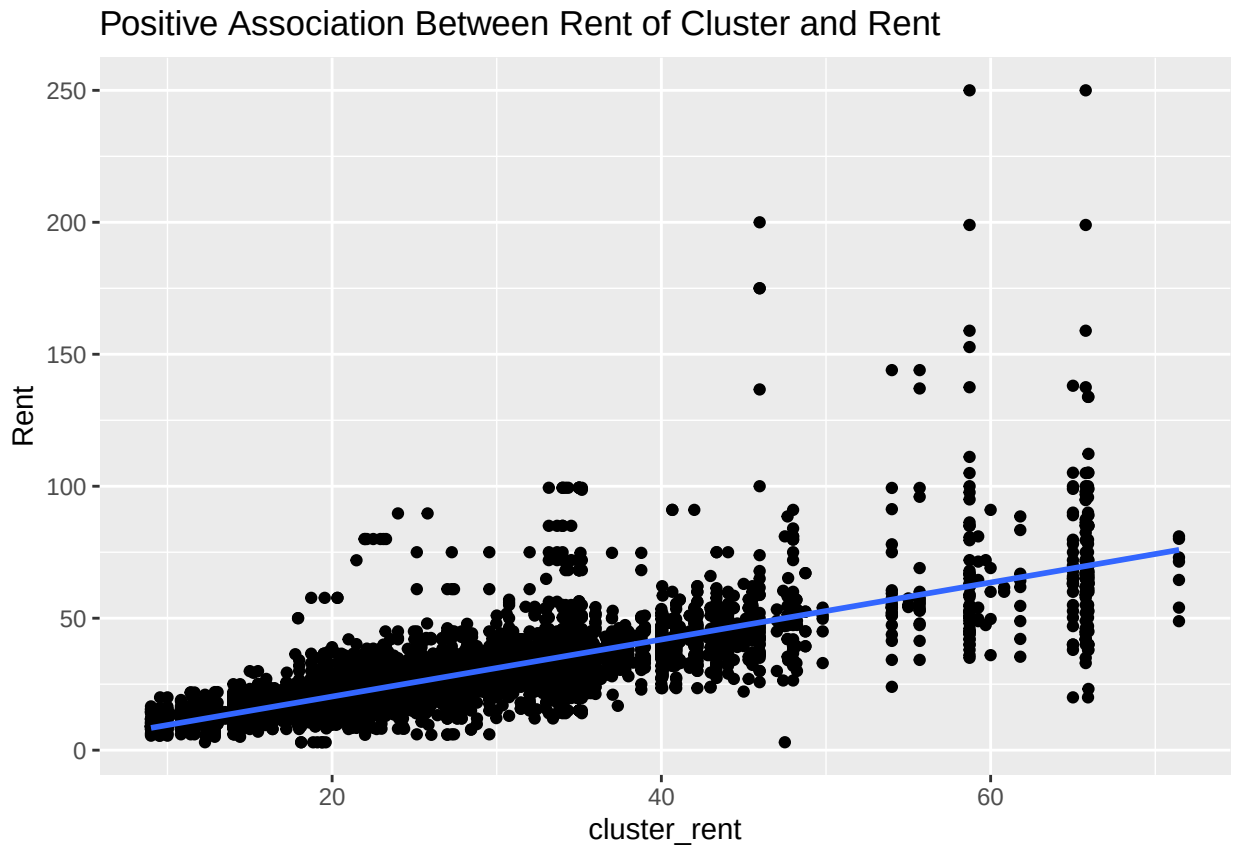
Yea so class a buildings have higher median rent. But within the classes, being green only has minimal impact on median rent. Class is a dominant factor.

```
## `summarise()` has grouped output by 'green_rating'. You can override using the `.groups` argument.
```



This is a strong relationship! Location matters!

```
## `geom_smooth()` using formula 'y ~ x'
```



So now we should just run a model, in an attempt to look at each affect independently.

```
##
## Call:
## lm(formula = Rent ~ ., data = model_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -53.869  -3.596  -0.531   2.497 174.533
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -7.716e+00  9.973e-01  -7.737 1.14e-14 ***
## size              6.686e-06  6.559e-07  10.193  < 2e-16 ***
## empl_gr           6.069e-02  1.693e-02   3.585 0.000340 ***
## leasing_rate      8.877e-03  5.320e-03   1.669 0.095196 .
## stories          -3.622e-02  1.617e-02  -2.240 0.025149 *
## age              -1.272e-02  4.713e-03  -2.698 0.006987 **
## renovated        -2.201e-01  2.566e-01  -0.858 0.390920
## class_a           2.854e+00  4.379e-01   6.518 7.58e-11 ***
## class_b           1.179e+00  3.428e-01   3.439 0.000587 ***
## LEED              1.901e+00  3.584e+00   0.530 0.595837
## Energystar       -4.444e-02  3.819e+00  -0.012 0.990715
## green_rating      5.536e-01  3.840e+00   0.144 0.885375
## net              -2.537e+00  5.931e-01  -4.278 1.91e-05 ***
## amenities         6.043e-01  2.504e-01   2.414 0.015809 *
```

```
## cd_total_07       -1.266e-04  1.464e-04  -0.865 0.387164
## hd_total07         5.369e-04  8.947e-05   6.002 2.04e-09 ***
## Precipitation      4.391e-02  1.598e-02   2.748 0.006014 **
## Gas_Costs         -3.444e+02  7.614e+01  -4.523 6.18e-06 ***
## Electricity_Costs  1.938e+02  2.489e+01   7.785 7.87e-15 ***
## cluster_rent       1.008e+00  1.402e-02  71.938  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.418 on 7800 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.6121, Adjusted R-squared:  0.6111
## F-statistic: 647.7 on 19 and 7800 DF,  p-value: < 2.2e-16
```