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.

Data Cleaning:

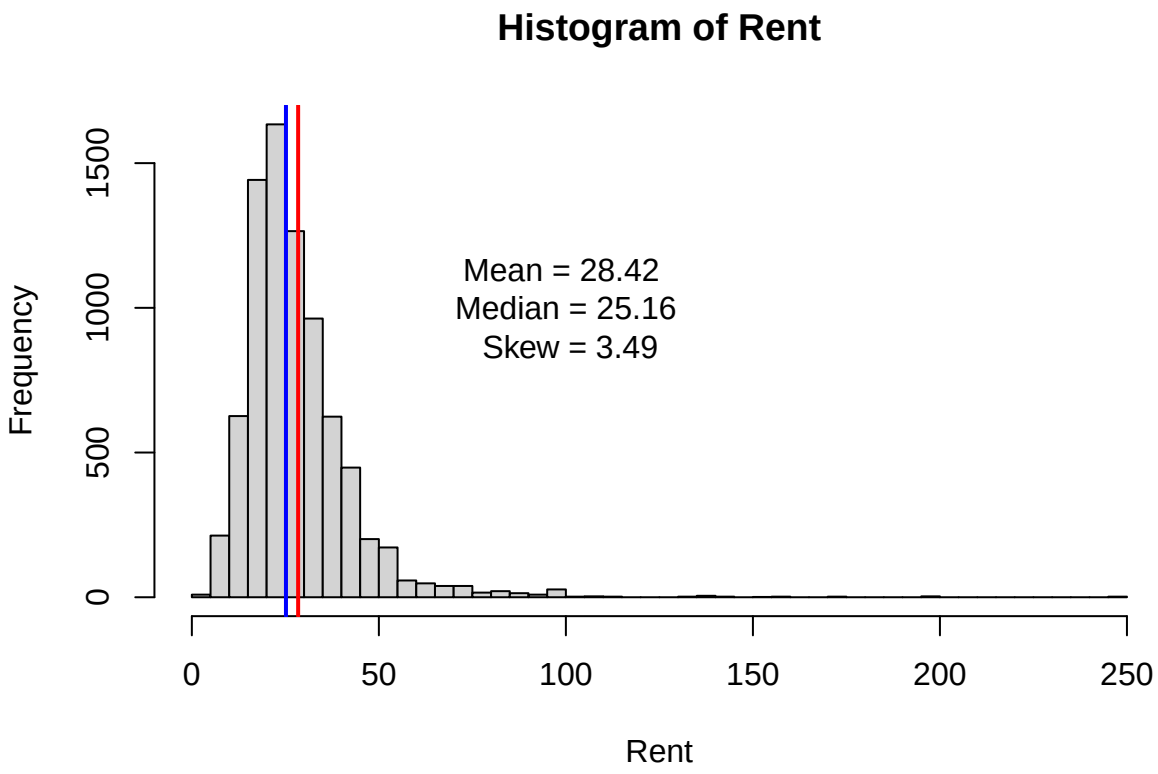
First, we can address the data cleaning aspect of the analysis. Are there any adjustments that should be made? The excel guru certainly felt that a certain amount of scrubbing should take place. He decided that because some of the buildings had low occupancy rates they should be excluded from the analysis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

ˆExplain the breakdown of his reasoning better here

Mean vs Median: Who's in Charge?

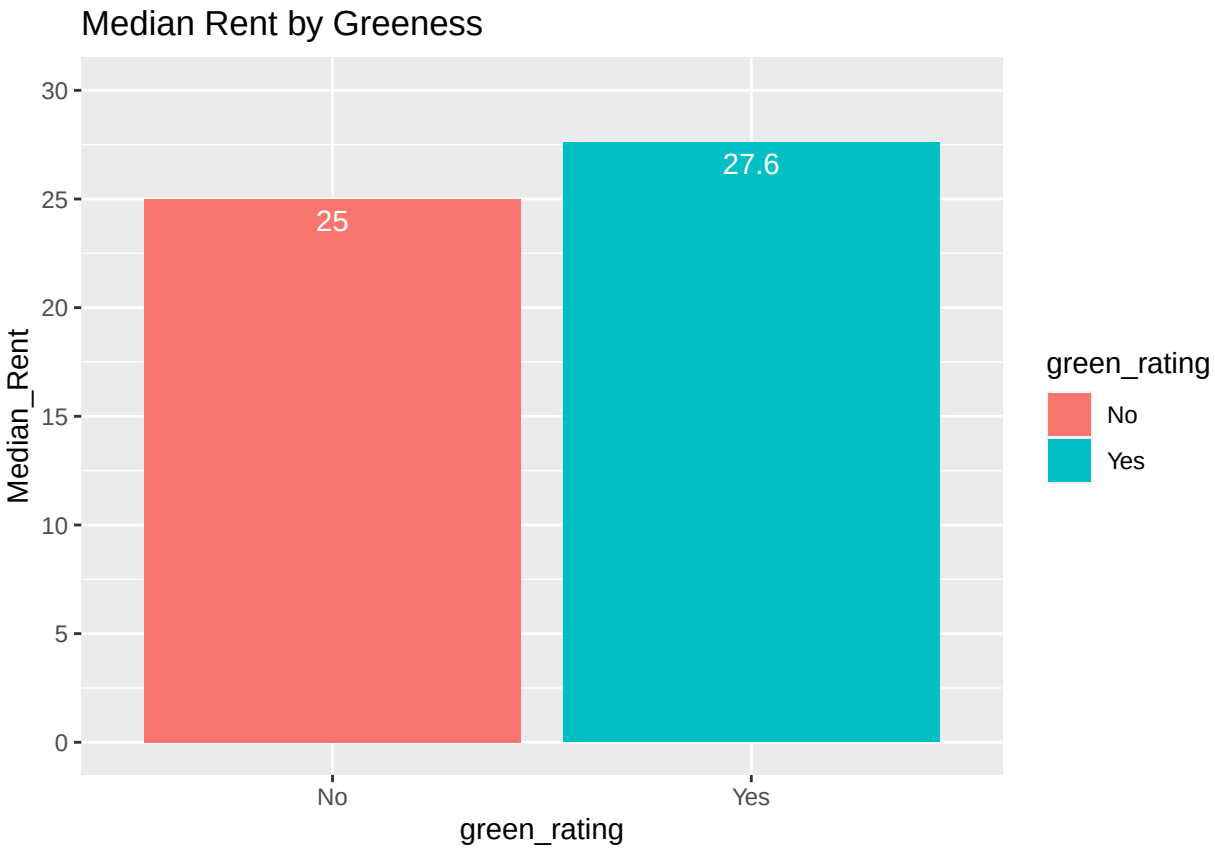
Note: For the purposes of presentation clarity, we have converted all categorical, binary variables from "1", "0" to "Yes","No".

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 that this newly constructed building will be 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.

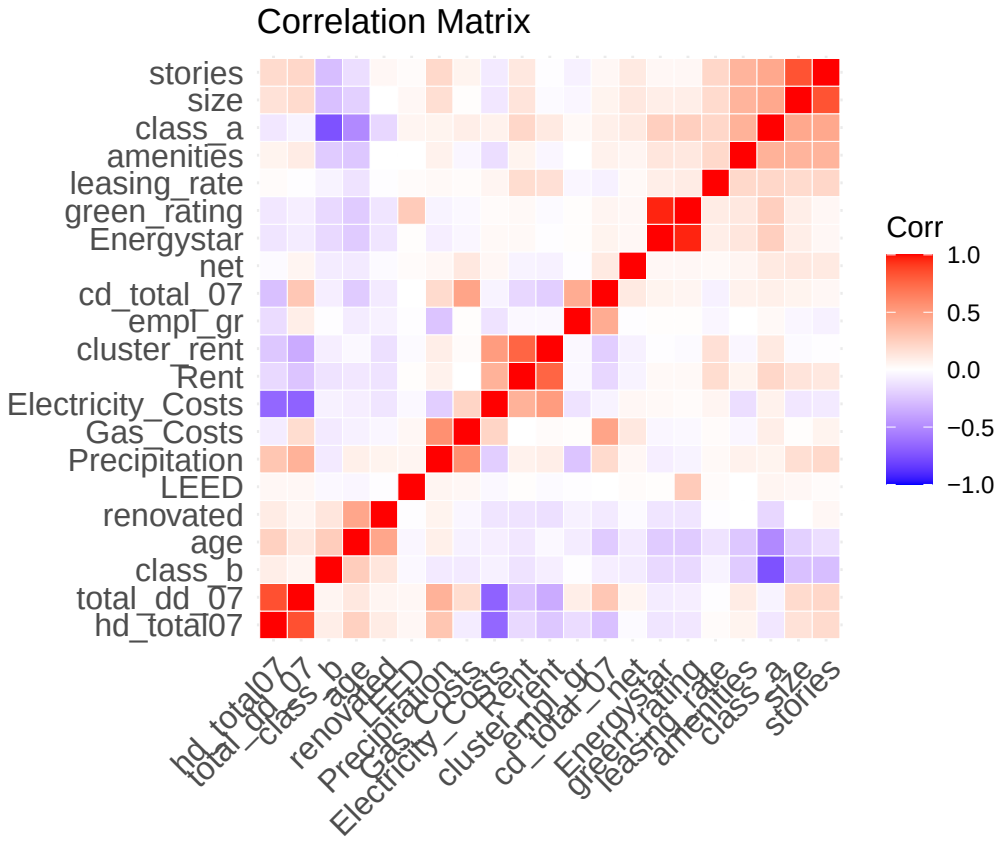


The median rent is certainly larger for green buildings vs non green buildings: \$27.60 per square foot for green buildings vs \$25.00 for non-green buildings. But this doesn't really tell the full story. It is too much a leap of faith to claim that this rental difference is due solely to the building's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

Confounders

We need to do some basic exploration of our data set beyond what we have already done with a goal of understanding how a buildings green rating is related to both rent and other variables. You might hypothesize that green buildings are simply associated with other factors that are really driving the difference in the median rental value. "Going Green" is a newer phenomenon so we might expect that the majority of buildings that have green certifications are newer buildings and that this newness is what drives their rent higher. Perhaps buildings that are built with green certification standard in mind are also built with higher quality overall and that this higher quality, indicated by building class, is what is determining rent dispersion. We must examine the data for the possibility of these confounders.

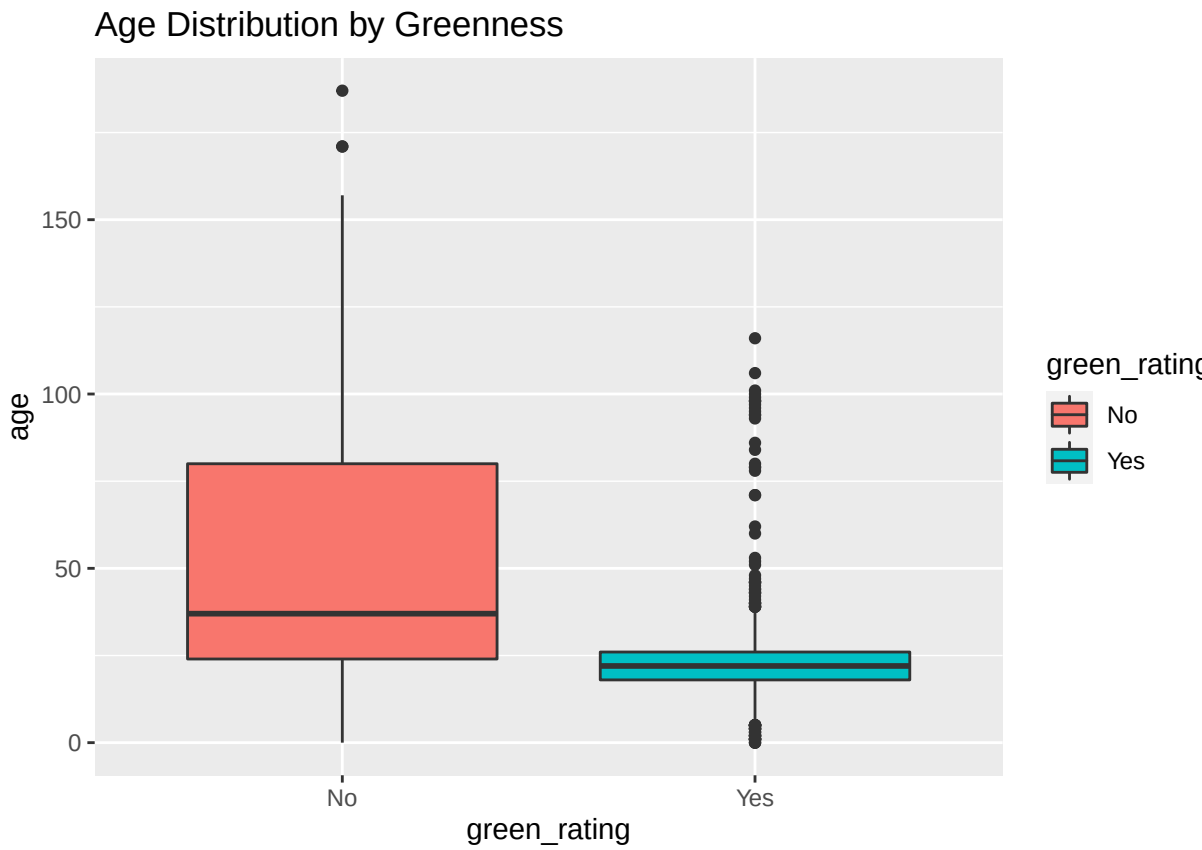
To get a feel for the potential interactions between variables in the data, we plot a correlation matrix.



There are lots of interesting relationships displayed here, but we would like to focus on a few that provide interesting information to the question at hand and test our hypothesized confounders: Age and Building Class.

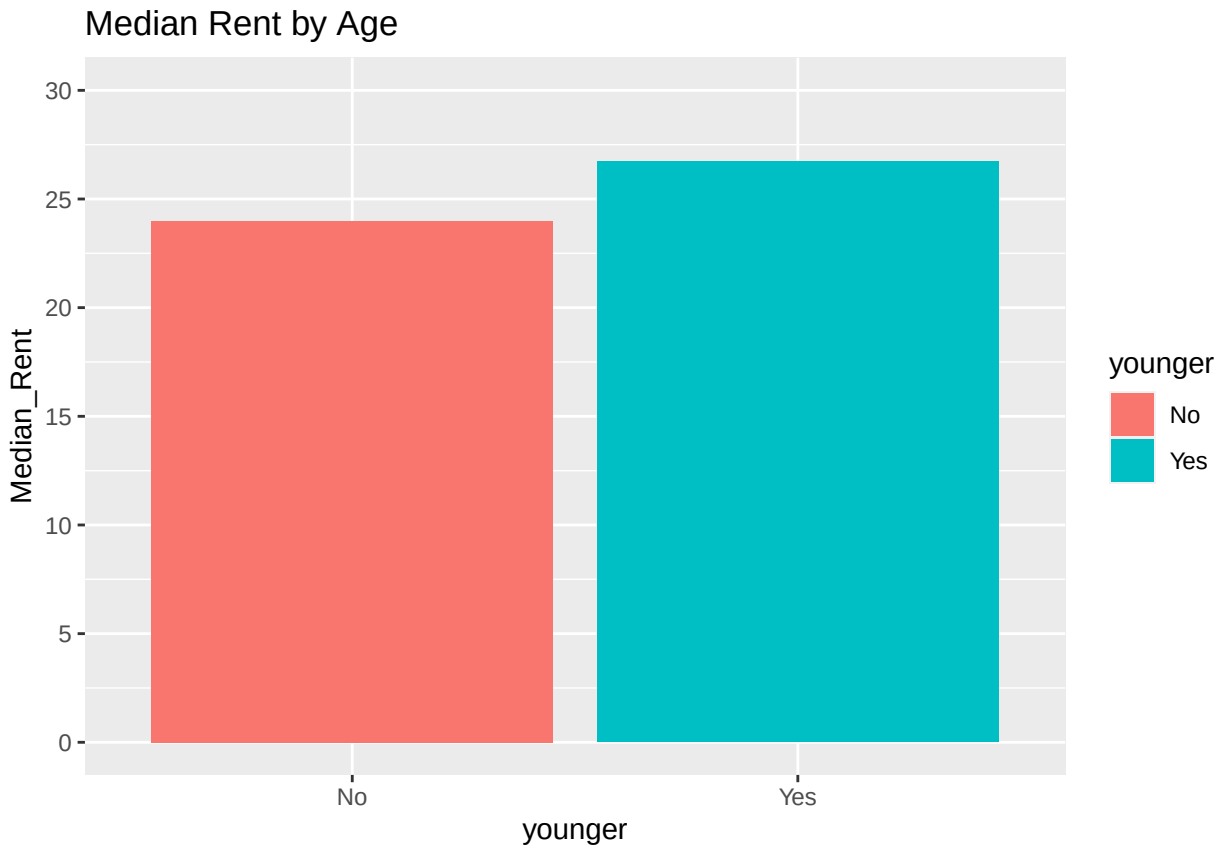
Age

First, examining the distribution of Age by green and non-green buildings, it's clear that green buildings do



tend to be younger.

We know that green buildings tend to be younger, but do younger buildings command a rent premium? To uncover a potential relationship here we create a new variable 'younger' that indicates if the building is below the median age of all buildings in our data set. This will allow us to control for the rent between young and old

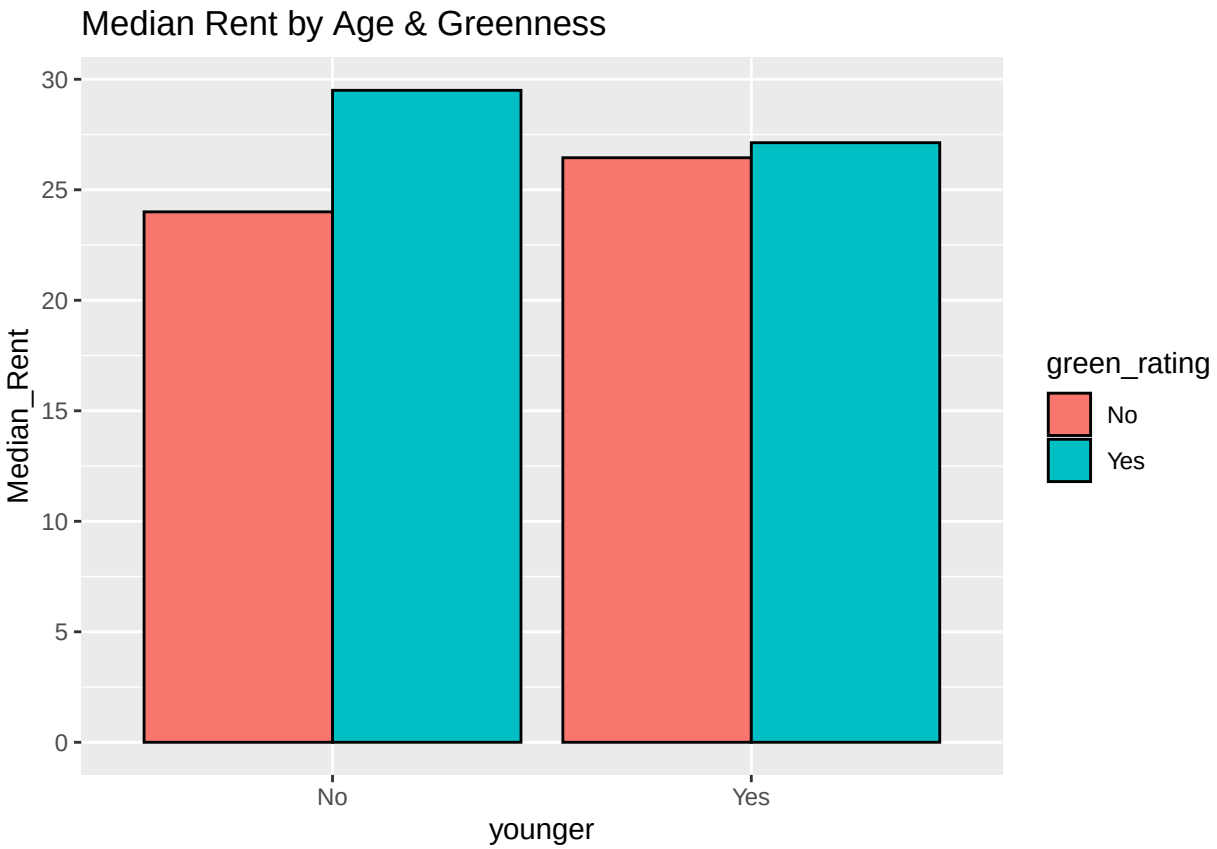


buildings:

We can see that younger buildings do have a higher median rent than older buildings. This should cast some doubt on the case for the green premium. We know green buildings tend to be younger and we know younger buildings tend to have a higher rent. How can we untangle this!?

We'll dig further by breaking down median rent by both green rating and age:

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

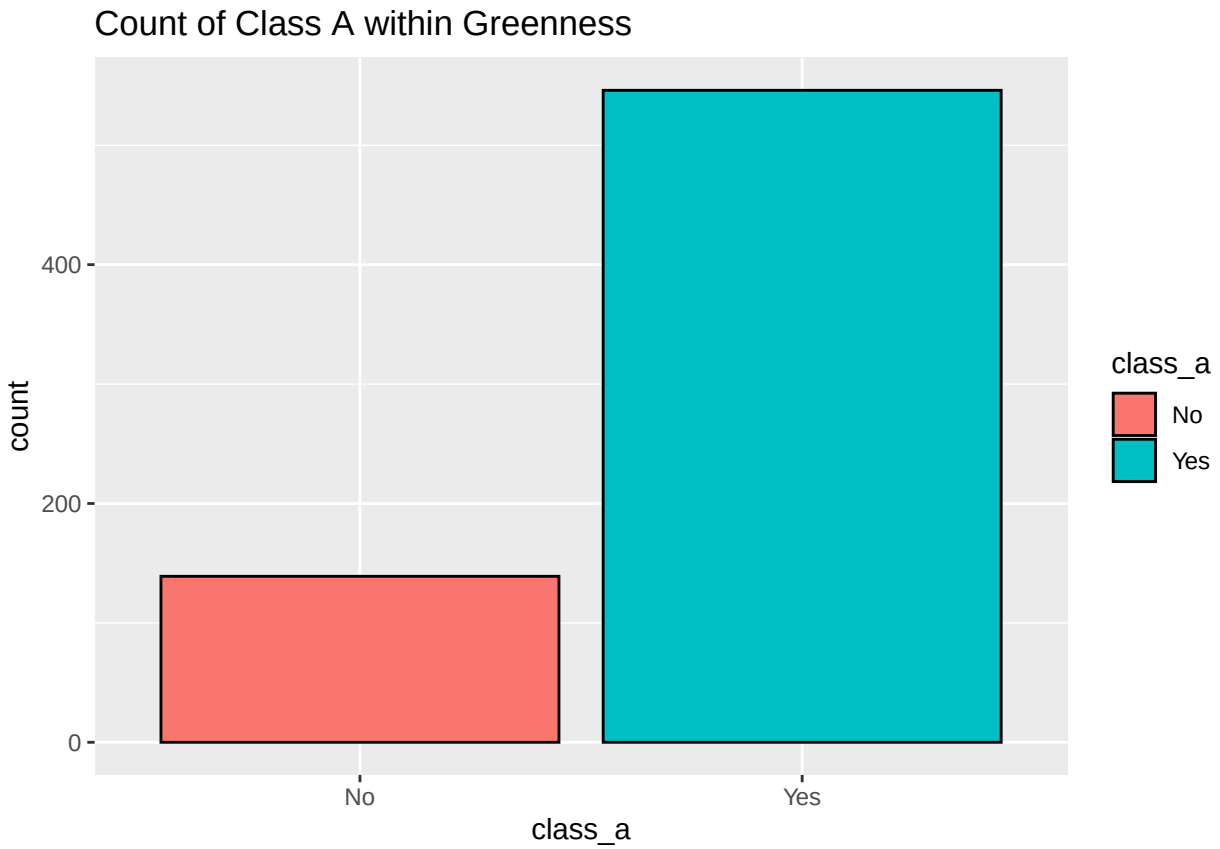


Interesting results. In both younger and older buildings, those that are green have a higher median rent than those that are not green. This lends a bit of credence to the green premium.

One intriguing point is that green buildings command a much higher premium in older buildings. For our purposes of determining the premium for the newly constructed building; however, the median rent differential among younger buildings is more relevant. In this category the premium is much smaller and we do not feel as confident in claiming the premium is significant and not due to chance or other confounders.

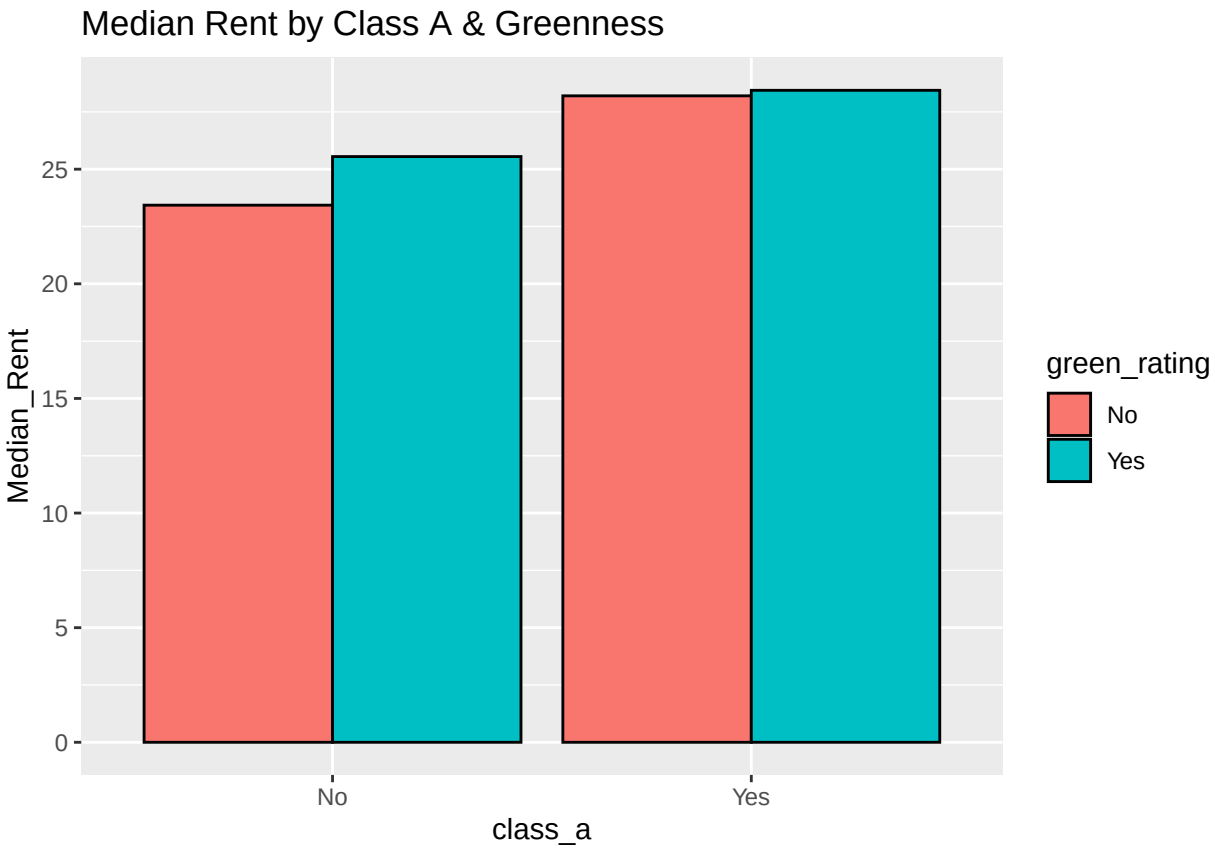
Class

What about class? We hypothesized that perhaps green buildings just happen to be built of a higher quality and thus higher class. We can see the evidence of this in the below plot. The majority of green buildings are also class A buildings.



Now we show a breakdown similar to Age, that allows us to control for being a class A building and examine the median of green vs non-green.

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

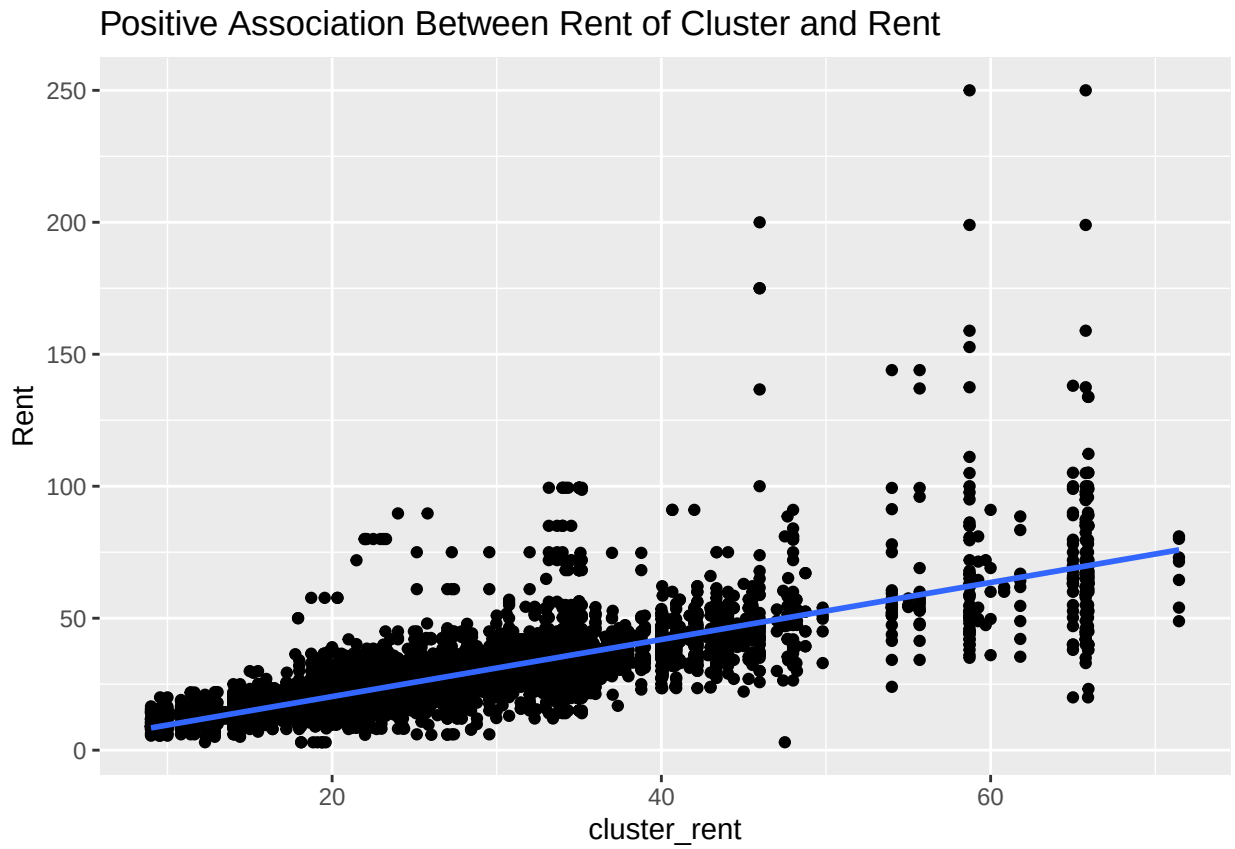


Class A is the winner in terms of rent, which does not come as a large surprise; we expect higher quality buildings to rent for more. There does appear to be a small advantage within each category for green buildings, but the dominant driver of rent here is the class.

Neighbor Rent

We also decided to investigate the relationship between the rent of other buildings in the local area with the rent of a particular building. Our correlation matrix gave us a strong positive association and we are always told that location is paramount when it comes to real estate. Below, you can see the Cluster Rent, Rent pairs plotted along with a fitted, signle-predictor, regression line. The association between local building rents and rent appears to be strong indeed although the variance does icnrease as the cluster rent increases.

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



Regression Model

As a final check on the analysis done so far, we run a multi-variable regression model with all of our independent variables. Our primary goal with this model is to validate and check the conclusions we have already made and ground our analysis with some numerical precision. The summary is depicted below:

```
##
## 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
```

Key takeaways from the model output. Age has a negative coefficient, class_a has a positive coefficient, and cluster rent has a positive coefficient. These match up with our previous analysis. Green Rating has a slightly positive coefficient of .5, but it is not statistically significant at the .05 level meaning that when we control for all of the factors in the data, the positive impact of green rating does not exhibit strong enough evidence that it truly exists.

Conclusion

Our final recommendation is: based on the data currently available, we do not recommend the developer should invest in the construction necessary to achieve a green certification. The data does not provide strong enough evidence to indicate that the green premium exists. The $ 5 million dollars could be more appropriately invested elsewhere.