Data Mining Exercise 3

Kylie Taylor, David Fraire, Larisa Barreto

4/4/2019

1) Determing rental price

To begin, we did a check that the variables make sense and are contained within reasonable bounds by investigating summary statistics.

Table 1: Table continues below

X	CS_PropertyID	cluster	size
Min. : 1	Min. : 1	Min. : 1.0	Min. : 1624
1st Qu.:1978	1st Qu.: 157407	1st Qu.: 272.0	1st Qu.: 50992
Median :3946	Median : 313253	Median : 476.0	Median : 128838
Mean :3954	Mean : 451337	Mean : 590.3	Mean : 234836
3rd Qu.:5933	3rd Qu.: 441062	3rd Qu.:1044.0	3rd Qu.: 294000
Max. :7894	Max. :6208103	Max. :1230.0	Max. :3781045

Table 2: Table continues below

empl_gr	Rent	leasing_rate	stories
Min. :-24.950	Min. : 2.98	Min. : 0.00	Min. : 1.00
1st Qu.: 1.740	1st Qu.: 19.50	1st Qu.: 77.89	1st Qu.: 4.00
Median : 1.970	Median : 25.20	Median : 89.54	Median : 10.00
Mean : 3.207	Mean : 28.42	Mean : 82.63	Mean : 13.59
3rd Qu.: 2.380	3rd Qu.: 34.18	3rd Qu.: 96.50	3rd Qu.: 19.00
Max. : 67.780	Max. :250.00	Max. :100.00	Max. :110.00

Table 3: Table continues below

age	renovated	class_a	class_b
Min. : 0.0	Min. :0.0000	Min. :0.0000	Min. :0.0000
1st Qu.: 23.0	1st Qu.:0.0000	1st Qu.:0.0000	1st Qu.:0.0000
Median : 34.0	Median :0.0000	Median :0.0000	Median :0.0000
Mean : 47.1	Mean :0.3788	Mean :0.3988	Mean :0.4596
3rd Qu.: 79.0	3rd Qu.:1.0000	3rd Qu.:1.0000	3rd Qu.:1.0000
Max. :171.0	Max. :1.0000	Max. :1.0000	Max. :1.0000

Table 4: Table continues below

LEED	Energystar	green_rating	net
Min. :0.000000	Min. :0.00000	Min. :0.00000	Min. :0.00000
1st Qu.:0.000000	1st Qu.:0.00000	1st Qu.:0.00000	1st Qu.:0.00000
Median :0.000000	Median :0.00000	Median :0.00000	Median :0.00000

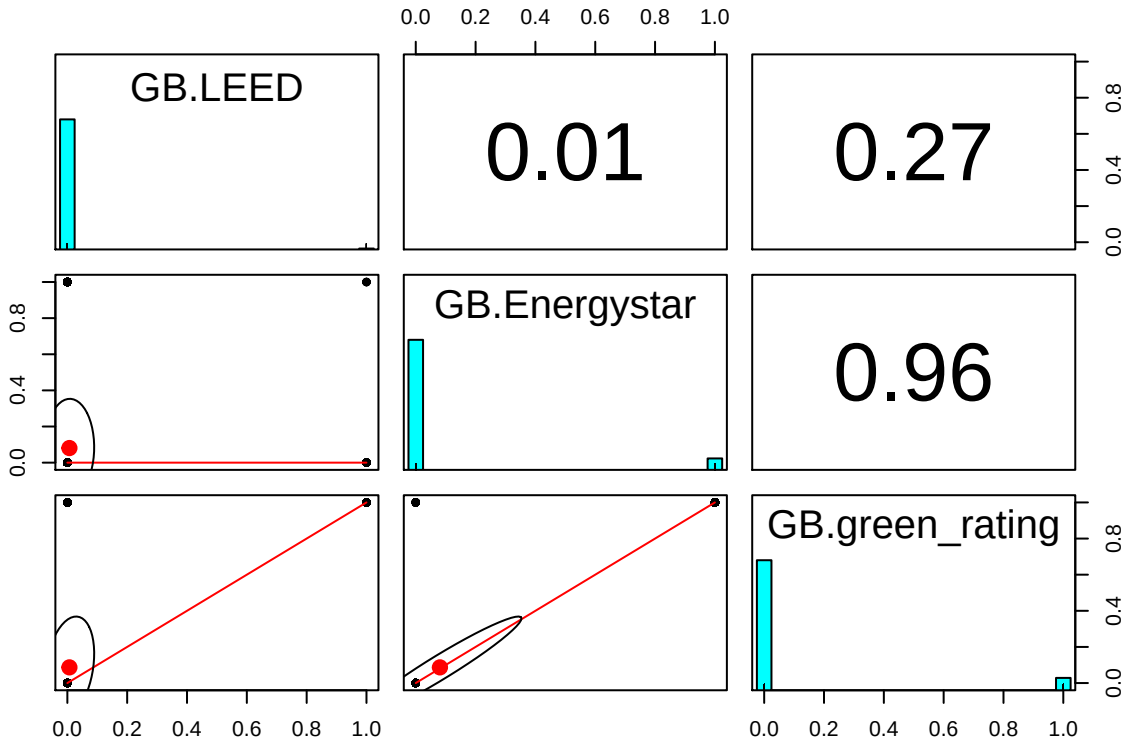
LEED	Energystar	green_rating	net
Mean :0.006905	Mean :0.08082	Mean :0.08683	Mean :0.03504
3rd Qu.:0.000000	3rd Qu.:0.00000	3rd Qu.:0.00000	3rd Qu.:0.00000
Max. :1.000000	Max. :1.00000	Max. :1.00000	Max. :1.00000

Table 5: Table continues below

amenities	cd_total_07	hd_total07	total_dd_07	Precipitation
Min. :0.0000	Min. : 39	Min. : 0	Min. :2103	Min. :10.46
1st Qu.:0.0000	1st Qu.: 684	1st Qu.:1419	1st Qu.:2869	1st Qu.:22.71
Median :1.0000	Median : 966	Median :2739	Median :4979	Median :23.16
Mean :0.5262	Mean :1232	Mean :3410	Mean :4642	Mean :30.84
3rd Qu.:1.0000	3rd Qu.:1620	3rd Qu.:4796	3rd Qu.:6363	3rd Qu.:42.57
Max. :1.0000	Max. :5240	Max. :7200	Max. :8244	Max. :57.00

Gas_Costs	Electricity_Costs	cluster_rent
Min. :0.009487	Min. :0.01780	Min. : 9.00
1st Qu.:0.010296	1st Qu.:0.02330	1st Qu.:20.17
Median :0.010296	Median :0.03095	Median :25.14
Mean :0.011305	Mean :0.03084	Mean :27.51
3rd Qu.:0.011816	3rd Qu.:0.03781	3rd Qu.:34.09
Max. :0.028914	Max. :0.06280	Max. :71.44

Since all variables appear to behaving well, we continued with our analysis. We investigated whether to include LEED and Energystar as seperate variables or to merge them into a single "green certified" variable. Below is a correlation plot of the variables LEED, Energystar, and green rating. Energystar has 96% correlation with green rating and LEED only has 27% correlation with green rating.



Based off the findings above, we decided to merge the two variables into one dummy variable called "green certified", which equals one if the building is both LEED and/or Energystar rated, and 0 if neither. There are 7,141 buildings that are not certified, and 679 that are.

The goal of this exercise is to build the best predictive model for price of rent. While there are many ways to approach this probelm, we decided to use stepwise selection to have R assist in picking the most robust model. In order to implement any automatic selection methodologies, we need to start with a working model. This allows us to combine some human nuance and the processing power of a modern computer to find the "best" model. In this problem we will define the "best" model as the one that has the lowest AIC and lowest out-of sample RMSE. It would be ideal to use cross validation to calculate out-of-sample RMSE when using stepwise selection, but the calculations are too extensive, therefore AIC is used to approximate MSE_{out}. The exact equation is $MSE_{AIC} = MSE_{in}(1 + \frac{p}{n})$. Instead, we will use 10 fold cross validation to determine the RMSE of the "best" model that stepwise selection suggests.

The steps taken for finding the best predictive model are:

1) Define a baseline model as

$$Rent_i = \beta_0 + green\_rating_i\beta_1 + ... + e_i$$

2) Use stepwise selection from the baseline to find model with lowest AIC
3) Include all "green certification" and building identification interactions in a stepwise selection process
4) Define a train test split of the Green Buildings data set and take out of sample RMSE of the "best" model
5) Preform lasso regression as another model selection process
6) Determine which model has lowest AIC and $RMSE_{out}$, and if any interacted coefficients are significiant in that model
7) Report coefficient of "green certification", if rent is different for different greeen certified buildings, and any other findings

The baseline model we started with is

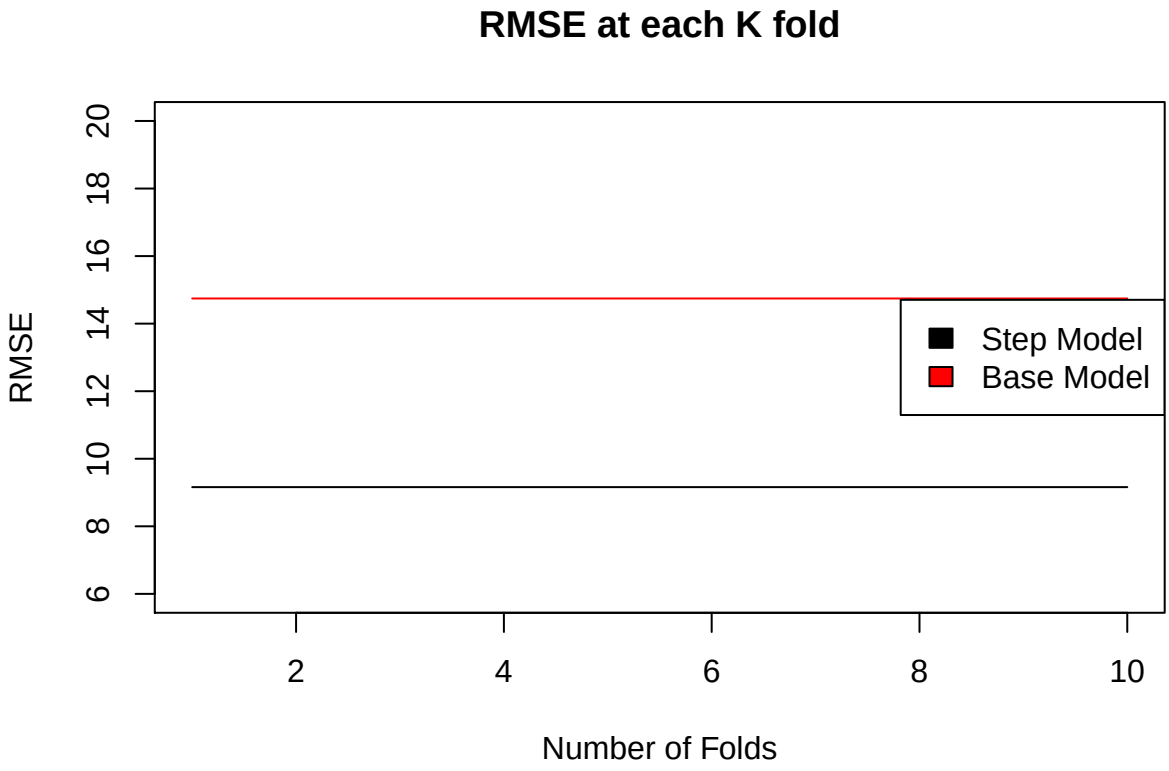
$$Rent_i = \beta_0 + green\_rating_i\beta_1 + class\_a_i\beta_2 + class\_b_i\beta_3 + e_i$$

We chose these variables since we need for green rating to be included in our final model, and we felt like the class of the building will effect the price of rent the most. The other variables were likely to be conditional on other factors of the building and surrounding buildings. For example, an old building may be very run down and cheap to rent, or historical, and high reputation and be expensive to rent. This is why we decided to begin with a simple working model.

We decided to skip over using a forwards and backwards selection process since we liked the flexibility stepwise selection offers. With forward selection, we consider only additions to our working model, and with backwards selection we only consider deletions from a full model. While both methods have their advantages, a stepwise process allows us to add and delete variables as the process suggests (or as the AIC suggests).

We decided to include all interaction terms as the scope of the stepwise selection process. Interaction terms may caputure a dynamic realtionship bewtween two input variables that has a significant effect on the rent price of a building. For example, the interaction berween cluster_rent (average rent in buildings local market) and size of the building proved to be a robust determinant in rental prices.

The stepwise selection identified 68 inputs that create the "best" model. The AIC for this model is 34,392.71. We conducted a 10 fold CV of the "best" model and found out of sample MSE converges to 83.23, or a RMSE of 9.123. We know that we have made an improvement from our baseline model, becuase the AIC is 57943.82, and CV out of sample MSE for the baseline is 217.39, or RMSE of 14.74. A plot of the RMSE values can be observed below.



We also ran a lasso regression to compare with the stepwise results. The lasso is a process for automated variable selection. The goal of a lasso regression is to minimize

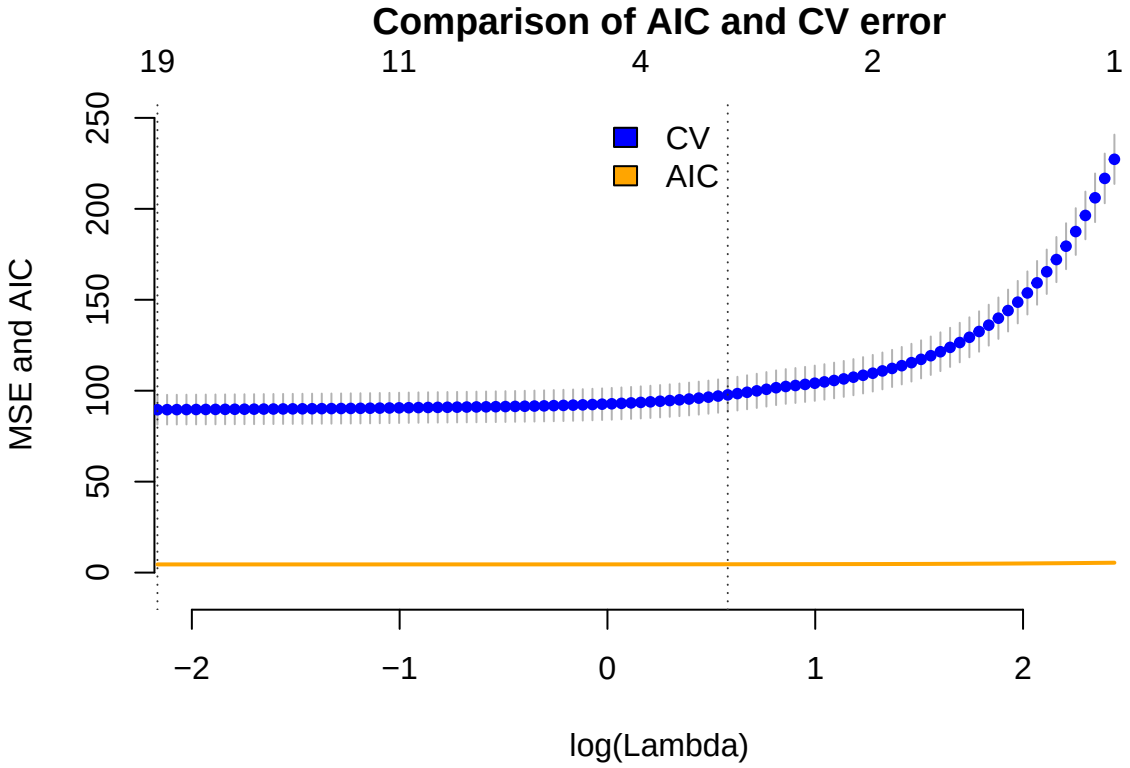
$$\frac{1}{n}dev(\beta) + \lambda pen(\beta)$$

where dev=deviance and is -2 times the log likelihood of the model.

This is done by using standard model selection tools, like AIC and RMSE, to find the best λ. The λ is the penalty weight, or cost function, that penalizes departure of the fitted β from 0. The β's that prevail for varying weights of lambda, suggest that including a given β is worth the penalty weight, λ, adding to the deviance. This means β's that prevail at the optimal λ are suggested to be significant in estimating the outcome of interest. A disclaimer about lasso is that selections are unstable and can vary greatly from different samples of the population.

The lasso suggests 19 main effect variables. The estimated β's that do not tend to deviate from zero are "CS_PropertyID", "cluster", "stories", "Energystar", "total_dd_07", and "Precipitation". This means that the lasso on this set of data suggests the above variables might not need to be included in a model. The best λ associated with these 19 variables is -2.166. The AIC at this λ is 36,370.89. The minimum deviance CV out-of-sample is associated with a $\log(\lambda) = 0.6719$. Based of these finidngs, we know we will be using the stepwise selected model, since it has a lower AIC.

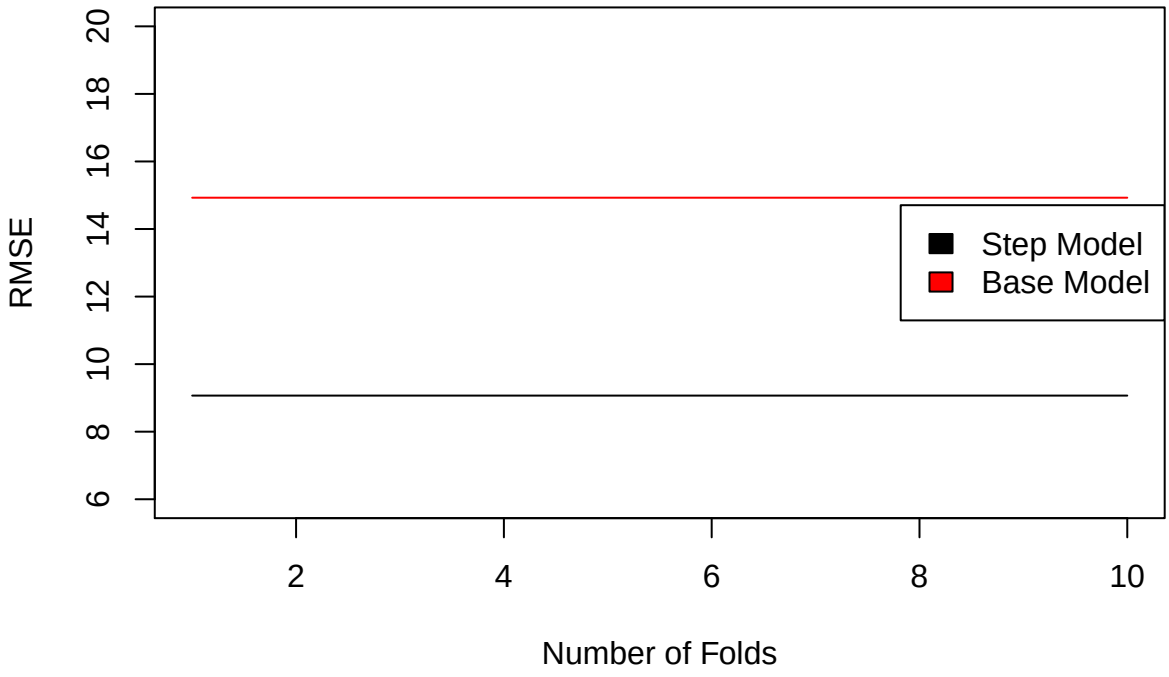
The plot below shows the cross-validated error rates and AIC at each respective λ value/penalty. It is easy to observe that the scaled AIC is much lower than the CV error, even at the "best" λ value. This reveals that the AIC under estimated the out-of-sample error rate at the optimal λ, in fact, for all λ's.



The last model we built was used to asses whether "green certification" effect is different for different buildings. We assumed that "different buildings" implied individual buildings, identified by the variable "CS_PropertyID". To begin, we difined a working model as

$$Rent_i = \beta_0 + greencertified\beta_1 + PropertyID\beta_2 + greencertified * PropertyID\beta_3 + e_i$$

We interacted certification and propertyID to identify if rent changes for each building if the building has certification. A stepwise selection process was implemented to determine the "best" model from the baseline. The stepwise selection suggest 69 variables in the model. This model has an AIC of 34,394.52 and out-of-sample MSE of 82.21, or RMSE of 9.07. This is an improvement from the baseline which has an AIC of 58,114.4, and an MSE of 222.69, or RMSE of 14.92. This can been observed in the plot below.



We used orthogonal machine learning to determine any idiosyncratic variation in rents by adjusting for individual buildings and green ratings. We found that the buildings with the ID's "5737391", "5768846", "5521767", "5625732", "4384675", "5697840", "5601727", "5513026", "4070629", and "5849472" had the highest sensitivity to changes in rent accounting for green rating status. On the other hand, we found that the buildings "5622696", "5335788", "5015583", "1380142", "6008486", "5713134", "5634370", "5622424", "5360564", and "5056278" were the top 10 least sensitive to changes in rent, accounting for green ratings.

Finally, to report results from the models we built, below are the outputs from the first model and the second model regressions (one with interaction of certification and ID). As the exercise asks, the first model reveals that the average change in rental income per square foot is expected to increase by $1.38 per square foot if the building has green certification, accounting for other features of the building. The exercise also asks to report if green certification is different for different buildings. While the estimate on Property ID is significant, the interaction between certification and ID is not significant. This reveals that changes in rental income per square foot for buildings with green certification is not likely to be *statistically* different for different buidlings.

Some interesting findings that the models revealed were statisically significant estimates for the class of the building (higher class buildings can expect higher rents), the cluster of buildings (certian clusters have consistently higher rents), and an interaction between size and electricity costs (for an average sized building, rent is expected to decrease as electricity costs increase).

```
##
## ================================================================================
##                                          Dependent variable:
##                                    -----------------------------------------------
##                                                    Rent
##                                           (1)                    (2)
## --------------------------------------------------------------------------------
## greencertified1                          1.514                  0.853
##                                         (0.980)                (1.008)
##
## CS_PropertyID                                                  0.00001***
```

```
##                                                                  (0.00000)
##
## cluster_rent                                    0.758***          0.949***
##                                                 (0.084)           (0.068)
##
## size                                           -0.00002***       -0.00001**
##                                                 (0.00001)         (0.00001)
##
## class_a                                         25.729***         19.352***
##                                                 (3.232)           (2.637)
##
## class_b                                         21.937***         17.254***
##                                                 (2.969)           (2.618)
##
## cd_total_07                                     0.001**           0.001**
##                                                 (0.0004)          (0.001)
##
## age                                              0.021             0.020
##                                                 (0.019)           (0.019)
##
## cluster                                         -0.007***         -0.008***
##                                                 (0.002)           (0.002)
##
## leasing_rate                                     -0.013            -0.010
##                                                 (0.018)           (0.016)
##
## net                                             -7.909**          -7.364**
##                                                 (3.757)           (3.499)
##
## Electricity_Costs                               -116.104          -26.389
##                                                 (88.909)          (78.567)
##
## empl_gr                                          -0.012           -0.071**
##                                                 (0.022)           (0.036)
##
## amenities                                       -2.297**          -2.264*
##                                                 (1.108)           (1.191)
##
## hd_total07                                      -0.0001           0.00004
##                                                 (0.0004)          (0.0004)
##
## stories                                                           -0.110
##                                                                   (0.071)
##
## greencertified1:CS_PropertyID                                     0.00000
##                                                                   (0.00000)
##
## cluster_rent:size                               0.00000***        0.00000***
##                                                 (0.00000)         (0.00000)
##
## size:cluster                                    0.000***          0.000***
##                                                 (0.000)           (0.000)
##
## cluster_rent:cluster                            0.0001*           0.00002
```

```
##                                    (0.00004)            (0.00003)
##
## size:leasing_rate                 0.00000***           0.00000***
##                                    (0.00000)            (0.00000)
##
## class_b:age                        -0.027***            -0.030***
##                                     (0.010)              (0.009)
##
## cluster_rent:leasing_rate           0.001**              0.001*
##                                     (0.001)              (0.001)
##
## CS_PropertyID:class_a                                  -0.00000***
##                                                         (0.00000)
##
## CS_PropertyID:age                                       -0.000*
##                                                         (0.000)
##
## size:Electricity_Costs              0.001***             0.0003***
##                                     (0.0001)             (0.0001)
##
## cluster:Electricity_Costs           0.161***             0.240***
##                                     (0.058)              (0.057)
##
## cd_total_07:net                      0.001*               0.001**
##                                     (0.001)              (0.001)
##
## cluster_rent:Electricity_Costs      4.214**
##                                     (1.794)
##
## class_a:empl_gr                      0.056*
##                                     (0.031)
##
## size:cd_total_07                     -0.000              -0.000***
##                                     (0.000)              (0.000)
##
## cluster_rent:age                   -0.002***            -0.002***
##                                     (0.001)              (0.0005)
##
## age:Electricity_Costs               1.280**              1.635***
##                                     (0.645)              (0.604)
##
## CS_PropertyID:class_b                                  -0.00000***
##                                                         (0.00000)
##
## greencertified1:amenities           -1.953**             -1.503*
##                                     (0.895)              (0.839)
##
## cluster_rent:class_a                -0.094*
##                                     (0.051)
##
## cluster_rent:class_b                -0.075*
##                                     (0.045)
##
## cluster:leasing_rate                -0.00002*            -0.00003*
```

```
##                                      (0.00001)                (0.00001)
##
## Electricity_Costs:amenities           52.541*                 105.371***
##                                      (31.687)                 (35.249)
##
## cluster_rent:amenities                                         -0.059**
##                                                                (0.028)
##
## CS_PropertyID:size                                             0.000***
##                                                                (0.000)
##
## CS_PropertyID:amenities                                        -0.00000*
##                                                                (0.00000)
##
## cluster:hd_total07                    0.00000*                 0.00000***
##                                      (0.00000)                (0.00000)
##
## cd_total_07:hd_total07             -0.00000***              -0.00000***
##                                      (0.00000)                (0.00000)
##
## Electricity_Costs:hd_total07          0.050***                 0.045***
##                                       (0.012)                  (0.012)
##
## cluster_rent:net                      -0.154**                 -0.123*
##                                       (0.068)                  (0.064)
##
## size:class_a                       -0.00002***              -0.00001***
##                                      (0.00000)                (0.00000)
##
## size:age                           -0.00000***              -0.00000***
##                                      (0.00000)                (0.00000)
##
## CS_PropertyID:empl_gr                                          0.00000*
##                                                               (0.00000)
##
## CS_PropertyID:Electricity_Costs                                -0.0001***
##                                                                (0.00002)
##
## CS_PropertyID:hd_total07                                       -0.000***
##                                                                (0.000)
##
## size:class_b                       -0.00001***              -0.00001***
##                                      (0.00000)                (0.00000)
##
## size:amenities                        0.00000**
##                                      (0.00000)
##
## class_b:amenities                      1.300**                  1.129**
##                                       (0.558)                  (0.525)
##
## cluster_rent:stories                                           -0.002
##                                                               (0.001)
##
## age:stories                                                    0.001**
```

```
##                                                                  (0.001)
##
## amenities:stories                                               0.082***
##                                                                 (0.024)
##
## empl_gr:stories                                                  0.003
##                                                                 (0.002)
##
## greencertified1:age                          0.028              0.034
##                                             (0.025)            (0.023)
##
## size:hd_total07                             0.000***
##                                             (0.000)
##
## class_a:hd_total07                          -0.002***          -0.001***
##                                              (0.0003)           (0.0003)
##
## cluster:stories                                                -0.0001**
##                                                                (0.00004)
##
## CS_PropertyID:cluster                                           -0.000*
##                                                                 (0.000)
##
## cluster:empl_gr                                                 0.0001**
##                                                                (0.00004)
##
## class_b:hd_total07                          -0.001***          -0.001***
##                                              (0.0003)           (0.0002)
##
## class_a:Electricity_Costs                  -384.146***        -292.094***
##                                             (77.467)           (64.187)
##
## class_b:Electricity_Costs                  -342.632***        -290.129***
##                                             (69.547)           (58.595)
##
## class_b:cd_total_07                         -0.001***          -0.001**
##                                              (0.0004)           (0.0002)
##
## class_a:cd_total_07                          -0.001**
##                                              (0.0004)
##
## net:Electricity_Costs                       173.597**          149.645*
##                                             (81.441)           (77.282)
##
## net:hd_total07                               0.001*             0.001
##                                             (0.0004)           (0.0004)
##
## age:cluster                                                    -0.00002*
##                                                                (0.00001)
##
## empl_gr:amenities                                               0.039
##                                                                 (0.027)
##
## cluster_rent:cd_total_07                                        -0.00003*
```

```
##                                                                    (0.00002)
##
## cd_total_07:stories                                                 0.00003*
##                                                                    (0.00002)
##
## hd_total07:stories                                                  0.00001
##                                                                    (0.00001)
##
## Constant                                        -2.228             -6.950**
##                                                (3.419)             (3.347)
##
## ------------------------------------------------------------------------------------
## Observations                                    7,038                7,038
## R2                                              0.640                0.664
## Adjusted R2                                     0.638                0.661
## Residual Std. Error                  9.215 (df = 6985)          8.592 (df = 6968)
## F Statistic                      239.246*** (df = 52; 6985) 199.850*** (df = 69; 6968)
## ====================================================================================
## Note:                                                  *p<0.1; **p<0.05; ***p<0.01
```

To conclude, determing factors that effect rent per square foot of buildings across the entire U.S. is no small feat. There are countless variables to account for and only a limited amount of information we have. Given the information we have, we tried to build the "best" model we can using stepwise selection and lasso regression. We have found that whether a building has green certification does result in higher expected rents, but is also highly dependent on other factors the building has. There is no perfect model, but given the tools and information we have, we can take out best shot at trying to determine important factors that can increase or decrease rent.

2)What causes what?

1. Why can't I just get data from a few different cities and run the regression of Crime on Police to understand how more cops in the streets affect crime?

The reason we can't just run Crime on Police and get an accurate understanding of how more cops affect crime is because such a regression would not only be lacking better specification, but also fail to establish causality, suffering from an extreme endogeneity issue. As we hear in the podcast, there are many factors we should take into account when trying to find true causality. Since we have different cities, it would be necessary to control for differences among them, including already-existing crime rates, police data, and other socioeconomic factors unique to each city. A model like this would also tend to be biased because cities with a high crime rate have the incentive to have large police forces. We need a more fine-tuned model to get the results we are looking for in this study.

2. How were the researchers from UPenn able to isolate this effect? Briefly describe their approach and discuss their result in the Table 2, from the researcher's paper.

The researchers from UPenn needed to collect data from a city where there are very specific fluctuations of police presence in the streets, unrelated to crime. They found the perfect example in D.C., a city where by law, extra police presence is dispatched for terror alert levels higher than orange. This is a very useful way to isolate the "Police" effect on crime, when they also controlled for other variables such as fluctuating tourist presence, measured by Metro ridership. Table 2 shows the basic regression that Klick and Tabarrok ran on daily D.C. crime totals against the terror alert level (1=high, Column 1) and then a second regression including daily Metro ridership (Column 2). For Column 1, the coefficient on the alert level is significant at the 5% level and indicates that on high-alert days, the total number of crimes decreases by an average of seven crimes per day, or approximately 6.6%. Column 2 the researchers ensure that the high-alert levels are

not being confounded with tourism levels by including a logged midday Metro ridership parameter in the regression. The findings were as follows: the coefficient on the high-alert parameter is slightly smaller, and that a 10% increase in Metro ridership only increases number of crimes by 1.7 crimes a day on average.

3. Why did they have to control for Metro ridership? What was that trying to capture?

To test if fewer visitors to the D.C. area could explain the results of the first regression, the researchers used Metro ridership data to control for such effect. Metro data suggested that there was a very small decrease in midday ridership on high alert days. Using this parameter as an instruent to capture tourism levels, the researchers were able to verify that the high-alert levels were not being confounded with visitor numbers.

4. Below I am showing you "Table 4" from the researchers' paper. Can you describe the model being estimated here? What is the conclusion?

The regression in Column 1 includes district fixed effects as well as the log of midday ridership. We see that during periods of high alert, crime in District 1 deceases by 2.62 crimes per day. We also see that crime in the other districts decreases by 0.571 crimes per day, but this number is not statistically significant.