Homework1

Andrew Bohl

8/7/2018

Probability Practice

A P(RC) = .3 P(Yes) = .65 P(Yes|RC) = .5 P(TC) = .7 P(No) = .35 P(No|RC) = .5

P(Yes) = P(Yes|RC)P(RC) + P(Yes|TC)P(TC) P(Yes|TC) = (P(Yes|RC)P(RC) - P(Yes)) / P(TC) P(Yes|TC) = (.5(.3) - .65) / .7 P(Yes|TC) = .71

B P(D+) = .000025 P(T+|D+) = .993 P(T+|D-) = .0001 P(D-) = .999975 P(T-|D+) = .007 P(T-|D-) = .9999

P(T+) = P(T+|D+)P(D+) + P(T-|D-)P(D-) = .993(.000025) + .0001(.999975) P(T+) = .00012

P(D+|T+) = P(T+|D+)P(D+)/P(T+) P(D+|T+) = .993(.000025)/.0012 P(D+|T+) = .198

If you test positive, there is still only a 20% chance that you have the disease. An issue with this test is if you test positive, there is a high chance that you will be going through taxing treatment plans when there is no need.

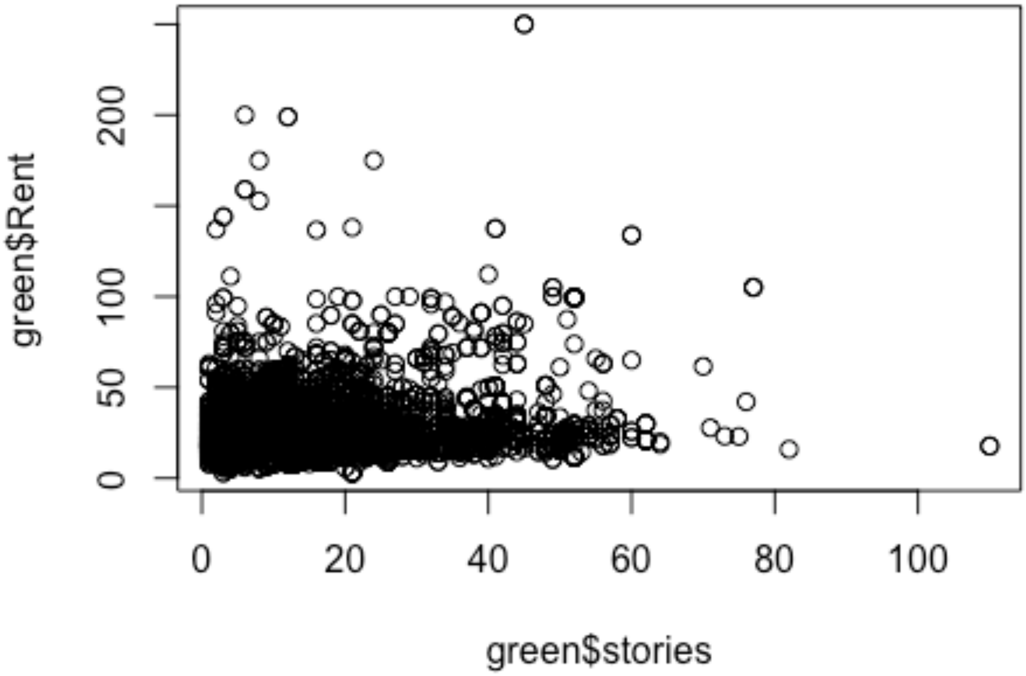
Exploratory analysis: green buildings

The first step that the analyst takes is to clean the data by removing any points where the leasing rate is below 10%. Looking further into this action and comparing the data, besides a decrease in the average number of stories present in a building the data is roughly the same. There are also only 215 occurrences of the low leasing rate, so taking these data points out is not necessary.

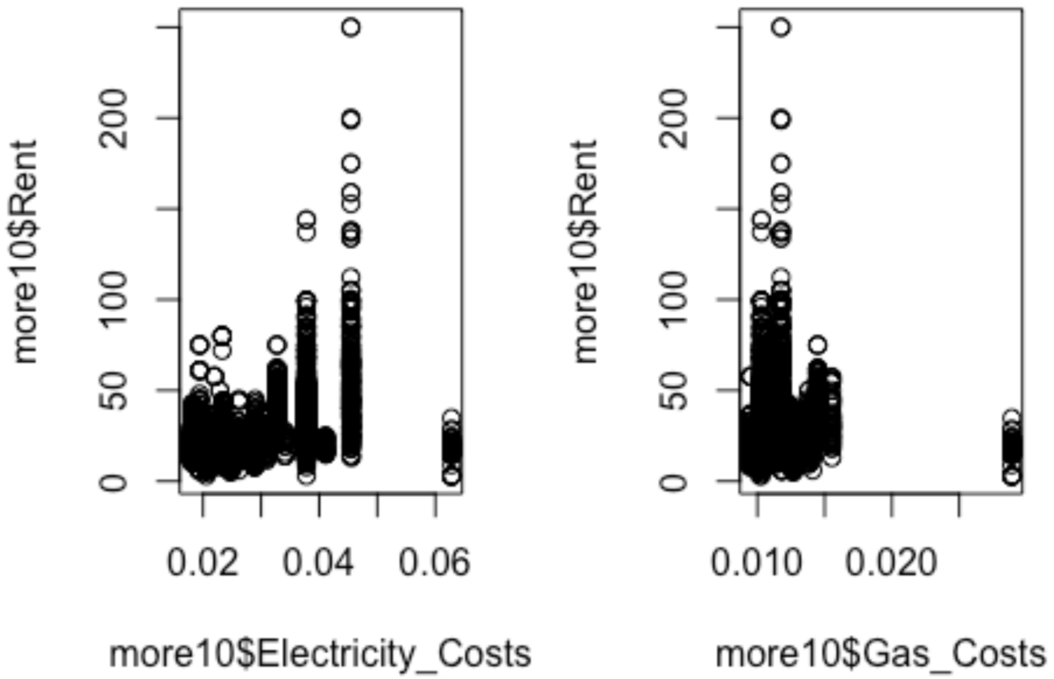
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.000   2.000   3.000   4.819   6.000  19.000

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00    4.00   10.00   13.83   20.00  110.00
```

On the top we have the summary for buildings with a leasing rate below 10%, and above 10% on the bottom.

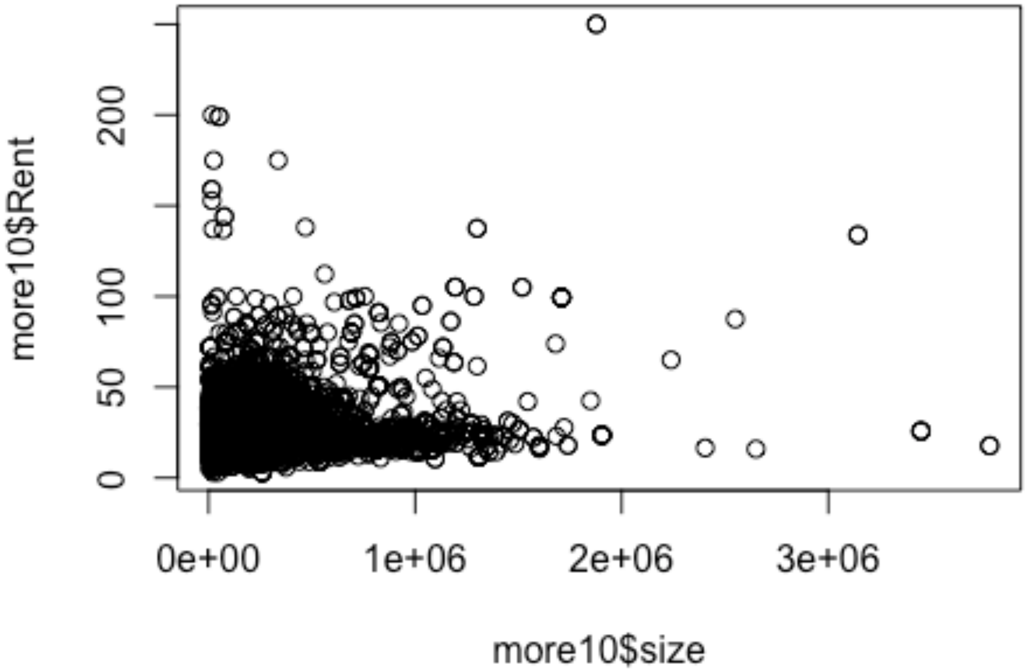
We want to start looking at the entire data set before delving into the green rating subset to see if any patterns exist regardless of the green rating. To do this I first looked at some pairwise graphs of rent compared to some features.



In the graph above, we can see a slight increase in the lower limit of rent as the number of stories increases. This could be from higher floors paying a premium to the lower and bringing the overall average up.

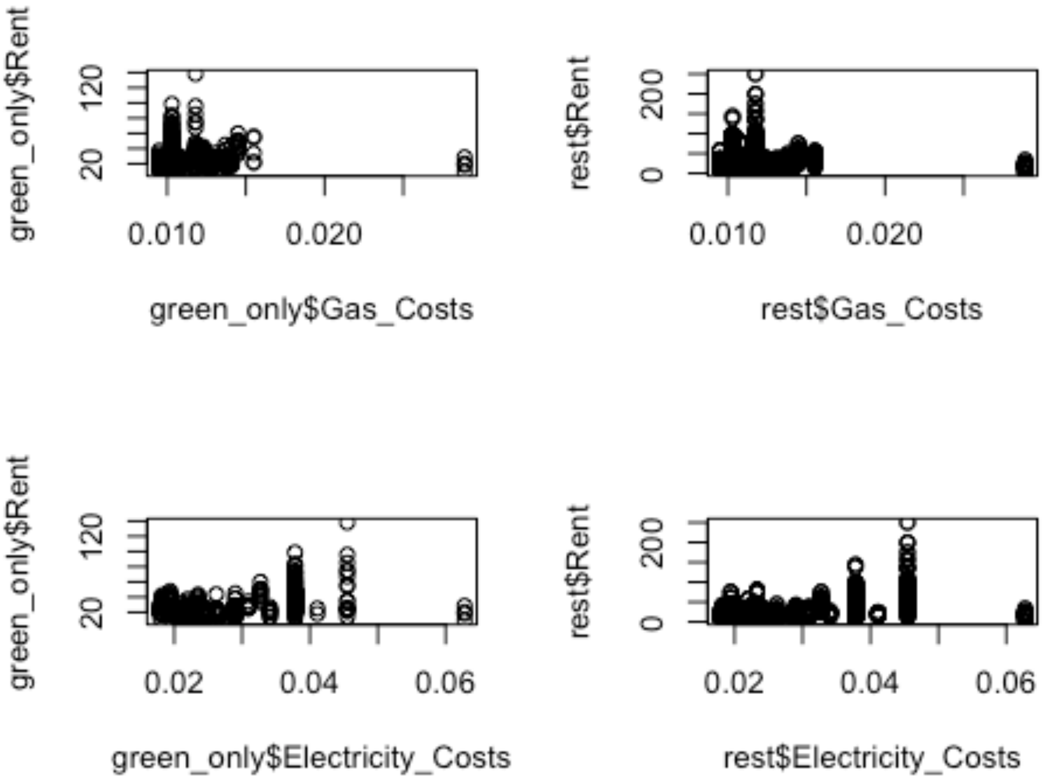


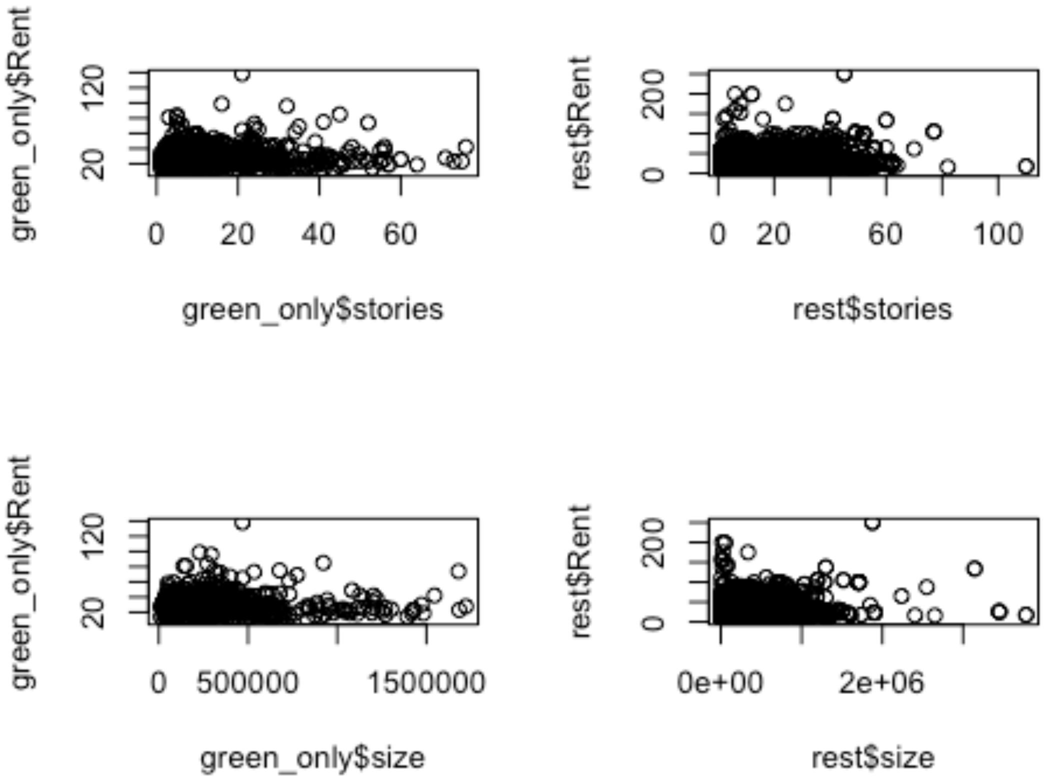
Looking at the utility costs versus rents, we can see a slight relationship that as utilities cost more in the region, the rent should also be higher.



The last plot in our preliminary research again shows a slight relationship in the rent cost as size increases. This is interesting as our rent is already measured in sq ft, however we still see increases in rent as the size of a building increases.

We now turn our attention to comparing buildings that were rated as green versus the rest of the buildings to see if going green really makes a financial difference. Looking at the same plots from before, we now compare the same variables on green and non-green buildings.


green_only$Rent
120
20
0.010
0.020
green_only$Gas_Costs
rest$Rent
200
0
0.010
0.020
rest$Gas_Costs
green_only$Rent
120
20
0.02
0.04
0.06
green_only$Electricity_Costs
rest$Rent
200
0
0.02
0.04
0.06
rest$Electricity_Costs



In comparing these graphs between the green buildings and the rest of the buildings in the data. We can't see too much of difference in rent price based on the green factor. The green only graphs look to just be a sample from the rest of the buildings. In conclusion, these variables seem to have an effect on the rent price independent of the green factor.

Finally, when addressing the mean and median of the rent price in comparing green and non-green buildings, I decided to look at rent of the clusters instead of the population. While using the median is a smart choice when looking at the full data set due to its robustness to outliers, I chose instead to look at the mean difference as we are taking rents from one location with similar features each time.

```
mean(green_only$Rent - green_only$cluster_rent)

## [1] 3.124401
```

When we compare the difference in rent on the cluster level, we see that green buildings do have a premium, however it is $.50 more than what we had originally thought. In this case we should still go ahead with the building, our only differences from the previous report is that we will be able to generate profits sooner.

In conclusion, the previous analysis was a good starting point and pointed out the key relationship in that green buildings can charge more for rent than non-greens.

However, the level of analysis could have been deeper looking at similar building styles and features to compare rents rather than a blanket statement about rent.

Bootstrapping

When we take a look at some statistics from the returns of all asset classes we see that Emerging-market equities has the greatest average return with only the second highest risk. In this case we defined the highest risk by which asset class had the lowest historical return for a given day. In addition to its high mean, Emerging-market equities also had maximum return way above the second closest asset class. From this it is safe to assume that EEM can produce extremely high returns, but also poses a high risk in comparison to the other 4 classes. Looking towards the safe side of investment, US Treasury bonds poses a smaller risk than the other 4 classes, again judging risk by looking at historical lows. The nice thing to note however, is that in all cases, our average return is positive showing that a buy and hold strategy should produce a positive return on investment if held for long enough.

```
##    ClCl.SPYa           ClCl.TLTa           ClCl.LQDa        
##  Min.   :-0.0984477   Min.   :-0.0504495   Min.   :-0.0911111  
##  1st Qu.:-0.0038636   1st Qu.:-0.0051997   1st Qu.:-0.0019083  
##  Median : 0.0006589   Median : 0.0005596   Median : 0.0004165  
##  Mean   : 0.0003981   Mean   : 0.0002788   Mean   : 0.0002095  
##  3rd Qu.: 0.0056254   3rd Qu.: 0.0057014   3rd Qu.: 0.0024660  
##  Max.   : 0.1451977   Max.   : 0.0516616   Max.   : 0.0976772  
##    ClCl.EEMa           ClCl.VNQa        
##  Min.   :-0.1616620   Min.   :-0.1951372  
##  1st Qu.:-0.0085338   1st Qu.:-0.0068896  
##  Median : 0.0008056   Median : 0.0006695  
##  Mean   : 0.0009814   Mean   : 0.0004209  
##  3rd Qu.: 0.0091897   3rd Qu.: 0.0077793  
##  Max.   : 1.8891250   Max.   : 0.1700654
```

In building my portfolios, I needed to look at addressing the wights put on to each asset class. By first looking at an even split of all classes, we can set up a control almost and compare our two portfolios.

In building an aggressive portfolio, my ideology was to shoot for the highest return we can get while slightly ignoring risk. To do so, I looked at the historical highs for the returns for each asset class. From this, I created assigned a weight proportional to the value of the high for any class to the sum of all highs. The point was to rely heavily on the top positive movers while trying to curb some risk by keeping some wealth in the more stable classes. The same basic method was used to create the safe method, only now we looked at minimizing loss instead of maximizing returns. The formulas used to receive the weights was similar on the back end, only we had to account for the fact we wanted smaller values to be higher in our final proportion.

Going along with what we saw in the summary from above, the heaviest weight in our aggressive model was placed on EEM, while the heaviest weight in our safe model was on TLT.

```
##           Even       Safe  Aggresive
##       935.7356   774.0094   2485.228
## 5% -6159.9734 -3873.7206 -11699.401
```

In the table above, the first row shows the expected return after 20 days of trading. In each case we can expect to see a positive return on average, we also see that there is a small difference between holding all classes evenly and creating a safer portfolio, however when looking at an aggressive portfolio, our returns are double the other options. The second row represents the value at risk with 95% confidence. In other words, only 5% of the time will you lose more than the amount listed in the second row depending on your investment strategy.

After comparing the data, the aggressive strategy should provide the highest returns, however it also comes with the much higher risk than other methods. While a safe strategy will most likely net the lowest returns of all strategies, but will also minimize the risk faced when investing.

Market segmentation

After performing principal component analysis on the data we can classify individuals to one or more market segments based on the amount of tweets that fall into any category. For each component, with the exception of the first (We ignore the first component in our calculations because all values are positive and will not give much insight into the data), we can pass in the amount of tweets to find a score for that component. The value of this score, if above 1 or below -1, will provide some possible segments that a user can be classified as. When we look over all components and all possible segments for any given user, we can reasonably assume that the user belongs to our calculated segment.

When we look at the individual components, we can see patterns arise in the data that show some correlation between segments. By grouping some of these together we can assign a user to one segment if they show strong interest in a correlated segment. For example, personal fitness tweets are often paired with health nutrition, in addition, art is often seen with TV and film. This is also seen in the data as the most common segments are personal fitness and health nutrition, with cooking and photo sharing.

For example, we can look at user 50.

```
SM[50,]
```

```
##            X chatter current_events travel photo_sharing
uncategorized
## 50 is65bq9kp       6              2     10             2
1
```

```
##    tv_film sports_fandom politics food family home_and_garden music
news
## 50       6             4        9    2      1               0     2
8
##    online_gaming shopping health_nutrition college_uni
sports_playing
## 50             2        0                1           1
1
##    cooking eco computers business outdoors crafts automotive art
religion
## 50       1   0         0        0        3      1          7   7
0
##    beauty parenting dating school personal_fitness fashion
small_business
## 50      0         2      0      0                1       0
0
##    spam adult
## 50    0     0
```

We can see this user tends to tweet about travel, politics, news, art, and automobiles. So manually we can safely assume that he fits into these markets. But lets try it using PCA.

```
pc1$x[50,]
```

```
##         PC1         PC2         PC3         PC4         PC5         PC6
##   2.4899390   0.2525174   5.0165367   1.1861316  -1.3200168  -1.5479417
##         PC7         PC8         PC9        PC10
##   1.5509688   5.0611744   1.4663369  -0.8822196
```

We will ignore PC1, PC2 and PC10 in this case as it does not meet our established threshold of 1. In addition, we saw very large values for PC3 and PC8 so we would want to weight these categories more heavily than the others. The top categories for PC8 include art, auto, and news with PC3 including travel, politics, news, and art. When we look at all possible categories with some weight based on the PC value, travel, politics, news, art, and automobiles would be the most prevalent, in line with what we can see looking manually at each tweet individually.