Excercise 3

Richard Marks

4/18/2020

`Question 1: Predictive Model Building`

The housing market seems very abarbitrary, there are some things that obviously affect the price of rent for a building, such as the age and the class. But there are many other things that to an outside observer seem like they should not affect the price at all. Using a Linear Regression we are looking to find the optimal model for the cost of rent, seeing how different factors and the interactions between the factors affect rents price. The other major thing this model will be doing is seeing if it is really worth the cost to make the building green, in essence if having a green rating increases the price of rent, and if so how much.

For this regression we first took out the Property ID as it is arbitrary and shouldnt be considered. Then we created a base model considering just the main effects of every single factor. After that we ran a stepwise model to find the optimal model for the price of rent(LML= lm(formula = Rent ~ cluster + size + empl_gr + leasing_rate + stories + age + renovated + class_a + class_b + LEED + green_rating + net + amenities + cd_total_07 + hd_total07 + Precipitation + Gas_Costs + Electricity_Costs + cluster_rent + size:cluster_rent + cluster:size + cluster:cluster_rent + size:Precipitation + stories:cluster_rent + size:leasing_rate + net:cd_total_07 + green_rating:amenities + LEED:cluster_rent + age:class_b + age:class_a + leasing_rate:cluster_rent + age:renovated + stories:amenities + class_b:Gas_Costs + size:cd_total_07 + class_b:amenities + cluster:leasing_rate + amenities:cluster_rent + age:cluster_rent + renovated:cluster_rent + age:Electricity_Costs + class_b:Precipitation + class_b:Electricity_Costs + class_a:Precipitation + class_a:Electricity_Costs + class_a:Gas_Costs + empl_gr:class_b + amenities:Electricity_Costs + hd_total07:cluster_rent + cluster:stories + net:cluster_rent + stories:cd_total_07 + renovated:hd_total07 + stories:renovated + size:renovated + stories:age + size:age + size:class_a + cluster:Electricity_Costs + cluster:hd_total07 + cd_total_07:hd_total07 + Electricity_Costs:cluster_rent + age:cd_total_07 + amenities:Gas_Costs + amenities:Precipitation + renovated:cd_total_07 + cluster:Gas_Costs + cluster:cd_total_07 + leasing_rate:LEED + size:stories, data = greenbuild)). While this formula looks complex there are a few interesting insights we can make, such the interactions of amenitites with different factors seems to be very important, as it is used multiple times in the model. Now that we have the optimal model it is time to see if being green certified leads to an increase in rent. To do this we just need to look at the main effect of green rating on rent, which is 2.13. This means that for being green certified leads to a 2.13 times rent increase. With this knowledge it seems to be a very good idea to build a green building, as you can charge double the rent than if it was a regular nongreen building

`Question 2: What Causes What`

`Part 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? ("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city.)`

The reason you cannot just run a simple regression on crime on police is because there are so many other factors that go into a cities crime rates then just the amount of officers present. You would be taking this incredibly complicated concept of crime rates that is affected by so many different factors, such as general

socioeconomic status of the citizens, if there are gangs, etc. and youre trying to pin it all on one single factor, that is just ludacrous. On another hand you cannot just run a regression, see a correlation between higher cops and lower crime, and say that more cops leads to less crime because correlation does not equal causation, becasuse of the complicated nature mentioned before in this answer.

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

What these researchers did was instead of just using the amount of police around to deter crime, they found another way for the city of Washington D.C. to have an increase in police officers, high terrorism alert days. On these days the police are not there to deter streetlevel crimes they are around to protect the city against possible terrorist threats. So the researchers looked to see if on these days if the crime level was lower and found that yes infact it is shown in the first column of table 2.The researchers then wanted to make sure that there is not less tourists in D.C. on high alert days, since toursit are the main targets of crime, to do this they added the D.C. METRO rider numbers to the equasion, which showed an equal number of tourists no matter the alert shown in column 2 of table 2. They found significance in police lowering the rate of crime.

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

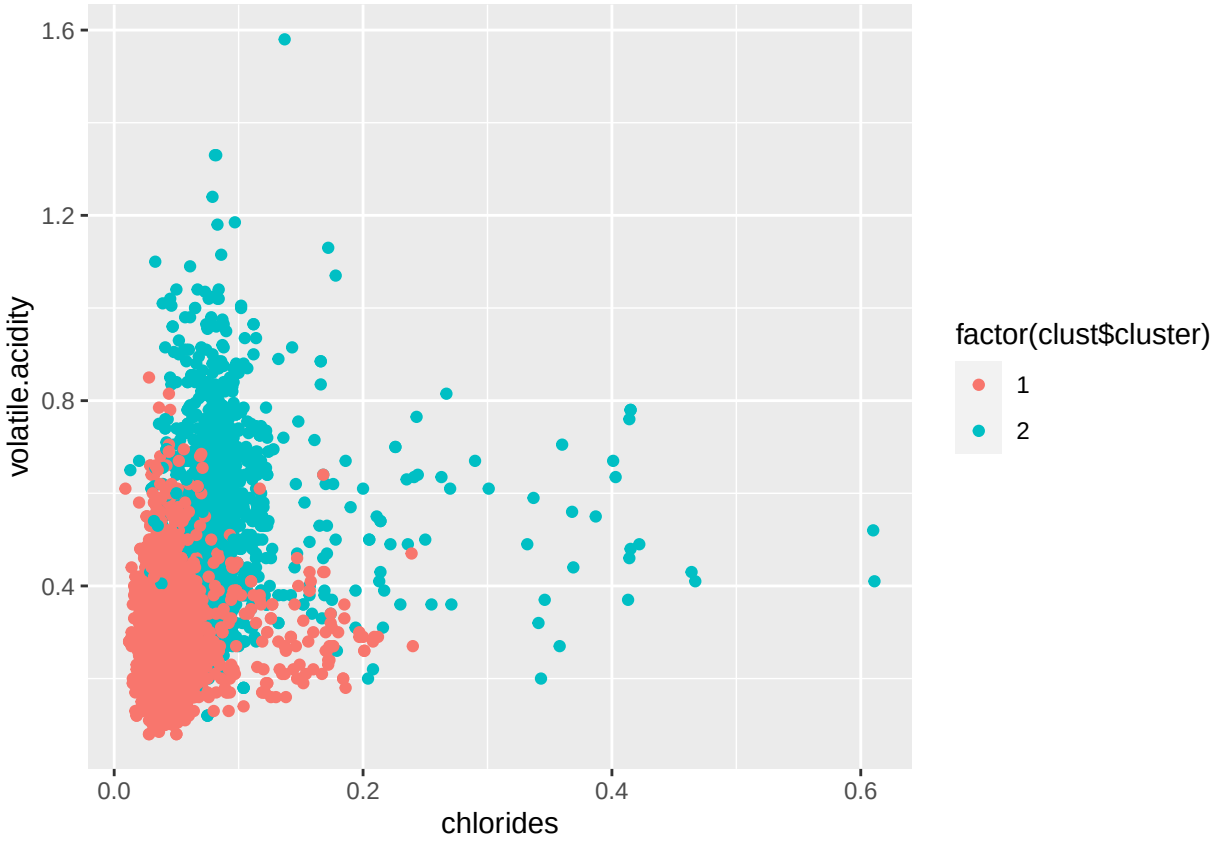
They had to control for Metro ridership because it shows how many tourists are in the city, since most travel via metro. If there was a significance decrease in metro ridership it would mean less tourist in the city, which leads to less targets for crime. They were trying to capture the amount of crime done on high threat days in washington D.C. and if they were less than low threat days becasuse of increased police numbers.

Part 4: Below I am showing you "Table 4" from the researchers' paper. Just focus on the first column of the table. Can you describe the model being estimated here? What is the conclusion?

The first column of table 4 shows all of the major parts of the model, the most important is the coefficent, which shows that High alert and high police prescence (High Alert x District 1) leads to crime having a zscore 2.621 standard deviations below the average crime rate. The next row (High Alert x Other Districts) shows that the high alert factor is not what lead to the decrease in crime, as it was only .5 standard deviations below the average crime. The thrid row (Log midday ridership) is to show that a higher level of people riding the Metro leads to an increase in crime. The conclusion is that a higher police prescence does in fact lead to a decrease in crime.

Question 3: Clustering & PCA

Clustering and PCA are two different ways to put in unsupervised data and get different information about classes out. For this task we are trying use the chemical composition of wine to tell if it is a red of a white, to start we will use a K-means clustering.



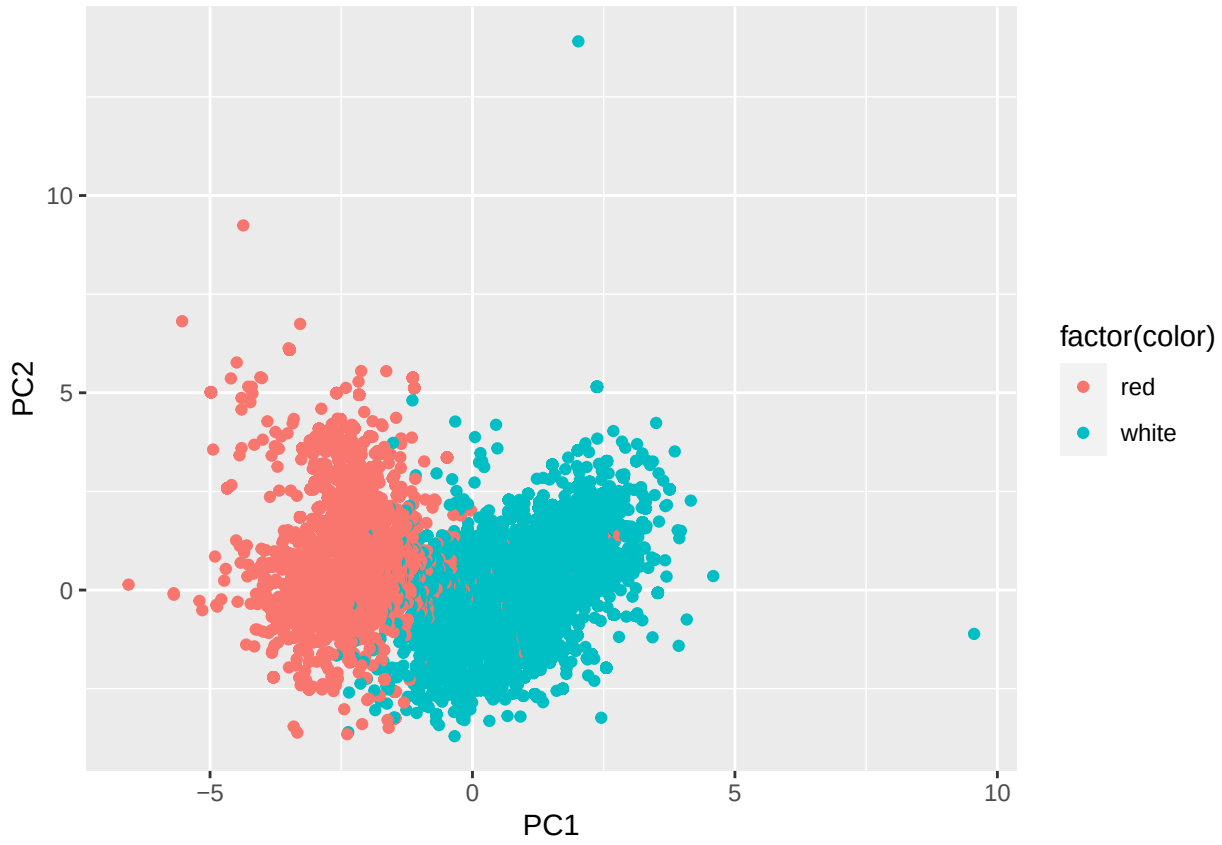

```
## [1] 29681.14 14931.51
```

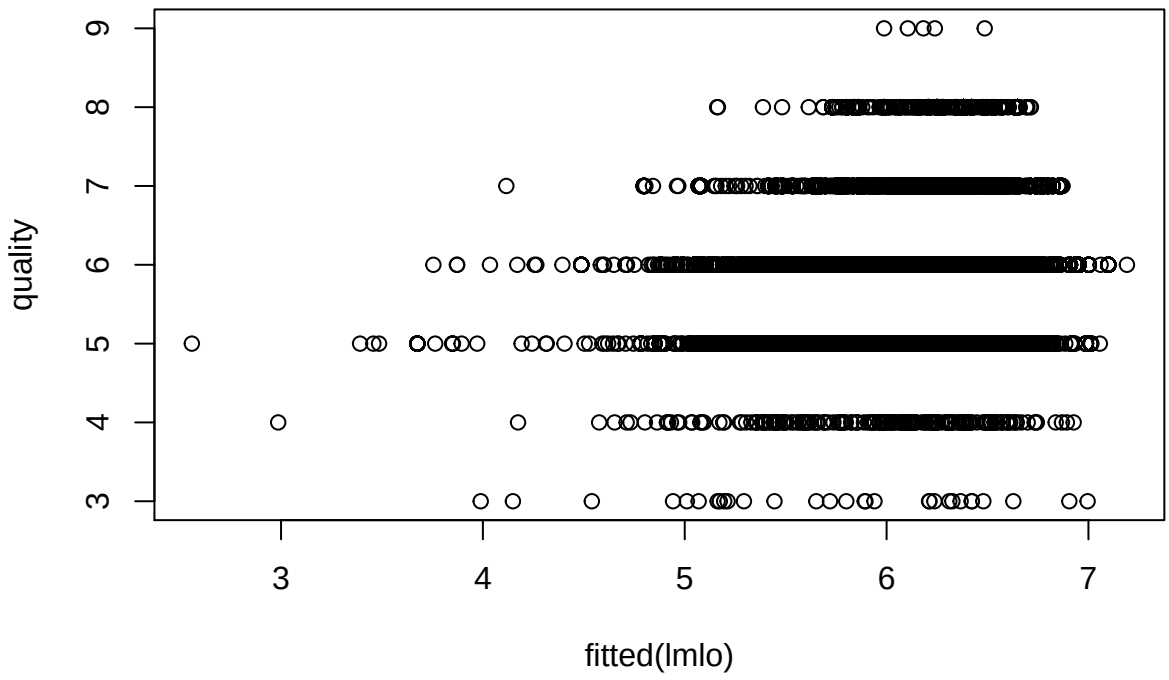
```
## [1] 44612.65
```

```
## [1] 13851.35
```

As shown in the graph above there are two very noticiable groups, with the red group being red wines and the blue being whites. While these groups are close togheter they are very distinct. However this method is not good at all at telling the quality of the wine. Next we will use PCA to try to group the wines, and see if the PCA can tell a good wine from a bad one.

```
##                         PC1   PC2   PC3   PC4
## fixed.acidity         -0.23  0.50 -0.30  0.12
## volatile.acidity      -0.41  0.10  0.36  0.08
## citric.acid            0.20  0.33 -0.49 -0.36
## residual.sugar         0.36  0.34  0.36  0.35
## chlorides             -0.29  0.33  0.11 -0.65
## free.sulfur.dioxide    0.46  0.02  0.21 -0.37
## total.sulfur.dioxide   0.51  0.03  0.17 -0.18
## density               -0.04  0.58  0.39  0.10
## pH                    -0.23 -0.28  0.40 -0.37
```


10
5
0
PC2
−5
0
5
10
PC1
factor(color)
red
white



As you can see from the plots above, PCA divided the wines color extremely well with very limited crossover. Also PCA is better at telling the difference between a good and bad wine ,normally putting it in at least 1 rank above or below its actual rank. With all of this evidence we can see that PCA is by far the surperior way to find the correct groupings from unsupervised data.

`Question 4: Market Segmentation`

In our modern era social media is king, millions of users log on daily, and no social media site is bigger and more popular than Twitter. For NutrientH20 its important to know what kind of people follow you, to get a better look into their intrests. To do this we categorized the tweets of all of NutrientH20's followers into 36 different very specific categories, of which we divided into 8 larger more general categories (News, Sports, Business, Relationships, School, Entertainment, Outdoors, Food.)

Below are the average amount of tweets per person per category

ENTERTAINMENT (10.64) NEWS (5.034) FOOD (5.97) SPORTS (3.7) SCHOOL (2.32) BUSINESS (.76) OUTDOORS (3.72) RELATIONSHIPS (3.59)

As you can see the three most tweeted about categories are Entertainment, Food, and News. Food is expected to be one of the highest, given that NutrientH20 is a beverage brand. Entertainment and News being higher is interesting, and gives us some insight into the dedicated consumer base of NutrientH20. These are the average tweets of consumers who follow NutrientH20, which means they are the most dedicated consumers. What this tells us is that our consumers most consistently like Entertainment, News and Food. For marketing it is very hard to make an ad campaign that goes off of current news, so that market is out, but we can work with entertainment, the single largest market we have. If we make adds based off of entertainment, such as famous TV shows, or we could try to partner with a famous youtuber. If we make tweets that are not overtly ads, such as "Lil Wayne must be thirsty after his performance on the Masked Singer!'. Tweets like these can get alot of retweets an likes, speading our brand and reaching new consumers.