Exercises

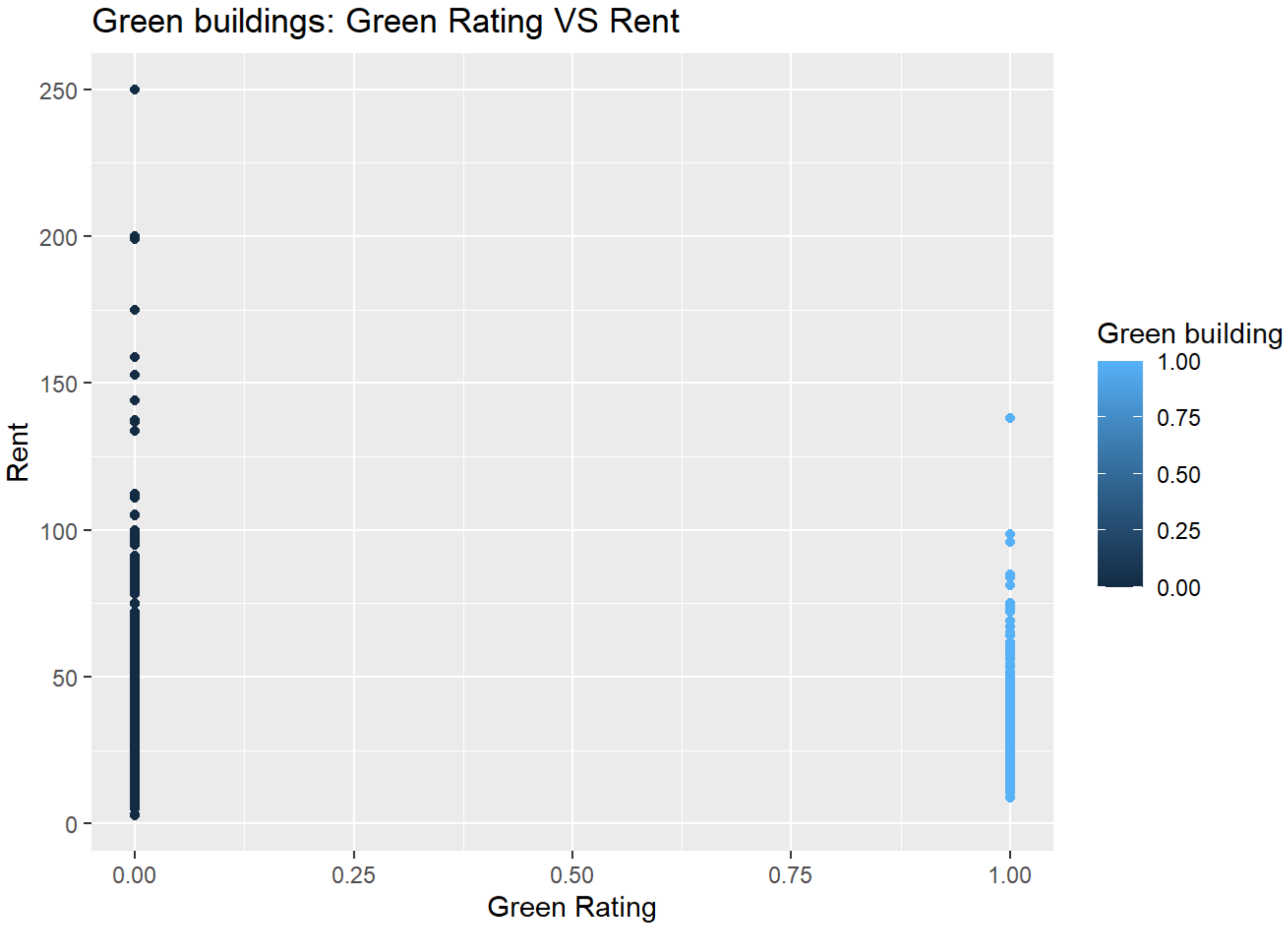
Jason Petri

8/5/2020

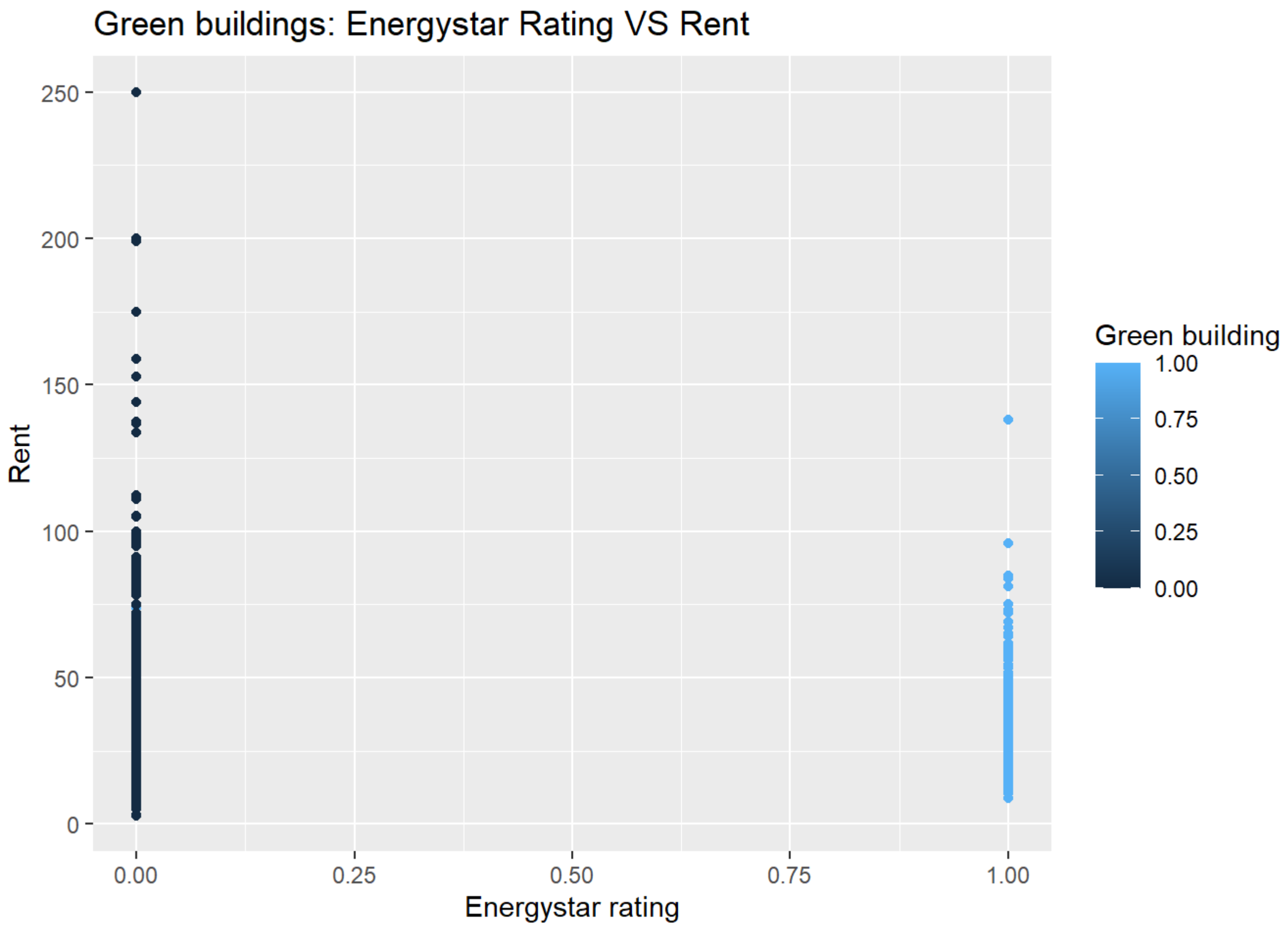
Visual story telling part 1: green buildings

```
##  CS_PropertyID        cluster            size            empl_gr       
##  Min.   :      1   Min.   :   1.0   Min.   :   1624   Min.   :-24.950  
##  1st Qu.: 157452   1st Qu.: 272.0   1st Qu.:  50891   1st Qu.:  1.740  
##  Median : 313253   Median : 476.0   Median : 128838   Median :  1.970  
##  Mean   : 453003   Mean   : 588.6   Mean   : 234638   Mean   :  3.207  
##  3rd Qu.: 441188   3rd Qu.:1044.0   3rd Qu.: 294212   3rd Qu.:  2.380  
##  Max.   :6208103   Max.   :1230.0   Max.   :3781045   Max.   : 67.780  
##                                                       NA's   :74       
##       Rent        leasing_rate       stories           age        
##  Min.   :  2.98   Min.   :  0.00   Min.   :  1.00   Min.   :  0.00  
##  1st Qu.: 19.50   1st Qu.: 77.85   1st Qu.:  4.00   1st Qu.: 23.00  
##  Median : 25.16   Median : 89.53   Median : 10.00   Median : 34.00  
##  Mean   : 28.42   Mean   : 82.61   Mean   : 13.58   Mean   : 47.24  
##  3rd Qu.: 34.18   3rd Qu.: 96.44   3rd Qu.: 19.00   3rd Qu.: 79.00  
##  Max.   :250.00   Max.   :100.00   Max.   :110.00   Max.   :187.00  
##                                                                     
##    renovated         class_a          class_b           LEED         
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.000000  
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.000000  
##  Mean   :0.3795   Mean   :0.3999   Mean   :0.4595   Mean   :0.006841  
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.000000  
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000000  
##                                                                       
##    Energystar       green_rating          net            amenities     
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.00000   Min.   :0.0000  
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000  
##  Median :0.00000   Median :0.00000   Median :0.00000   Median :1.0000  
##  Mean   :0.08082   Mean   :0.08677   Mean   :0.03471   Mean   :0.5266  
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000  
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.00000   Max.   :1.0000  
##                                                                         
##   cd_total_07     hd_total07    total_dd_07   Precipitation  
##  Min.   :  39   Min.   :   0   Min.   :2103   Min.   :10.46  
##  1st Qu.: 684   1st Qu.:1419   1st Qu.:2869   1st Qu.:22.71  
##  Median : 966   Median :2739   Median :4979   Median :23.16  
##  Mean   :1229   Mean   :3432   Mean   :4661   Mean   :31.08  
##  3rd Qu.:1620   3rd Qu.:4796   3rd Qu.:6413   3rd Qu.:43.89  
##  Max.   :5240   Max.   :7200   Max.   :8244   Max.   :58.02  
##                                                              
##    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.:20.00  
##  Median :0.010296   Median :0.03274   Median :25.14  
##  Mean   :0.011336   Mean   :0.03096   Mean   :27.50  
##  3rd Qu.:0.011816   3rd Qu.:0.03781   3rd Qu.:34.00  
##  Max.   :0.028914   Max.   :0.06280   Max.   :71.44  
##                                                       
```

Above are the summary statistics for the buildings data set. Rent will be a primary projection and we must declare the relationships between rent and other features of this data set. Rent ranges from a low end of $2.98 to $250.00. This is a large difference. There should be other factors for these properties that determine this wide variety.

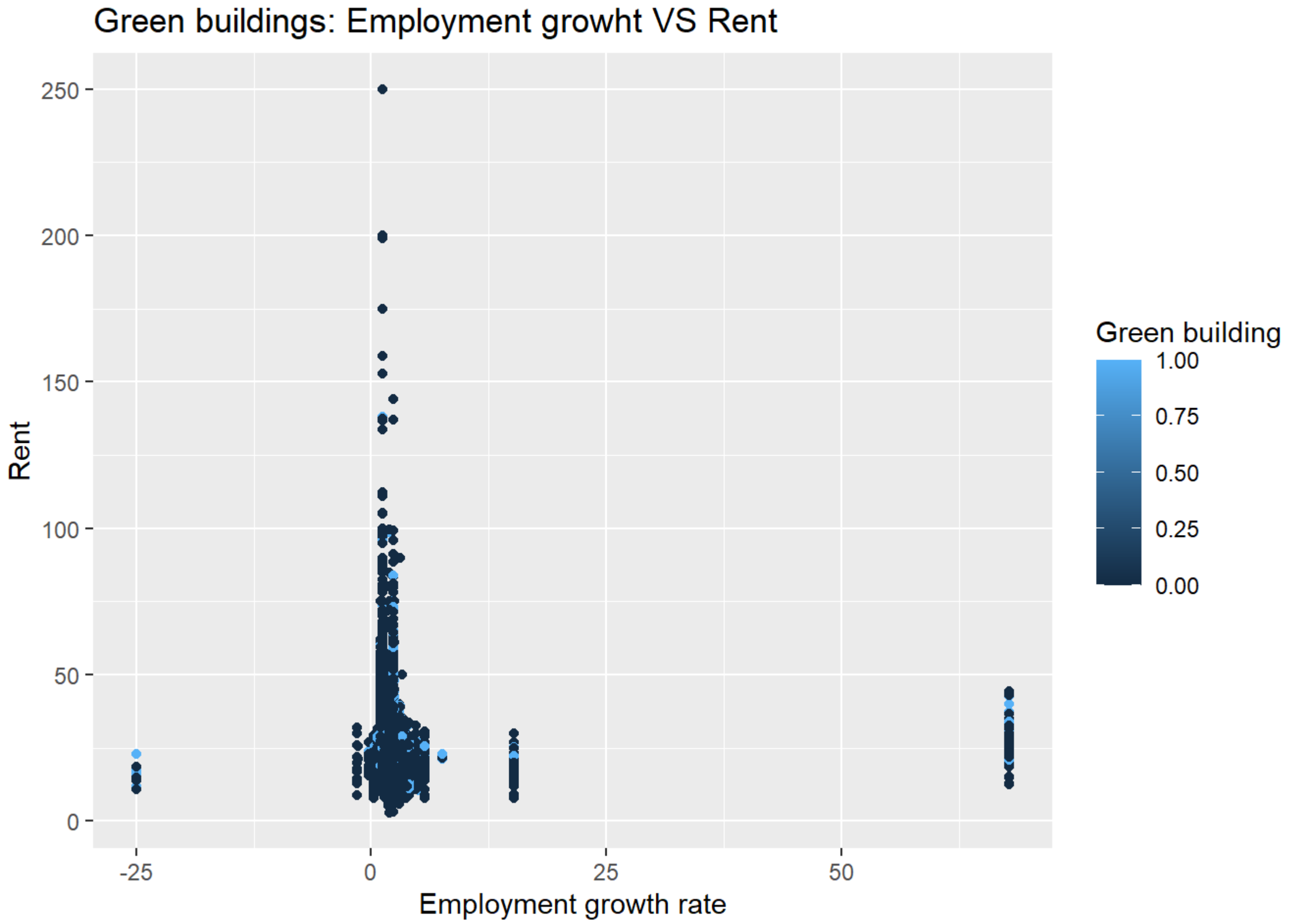


The preliminary analysis declared in the case begs the question: is there an apparent relationship between rent and the green rating. Above, it seems that the variance of rent charged is much lower than non-green buildings. Additionally, the average rent is situated lower for green rated buildings than normal buildings.



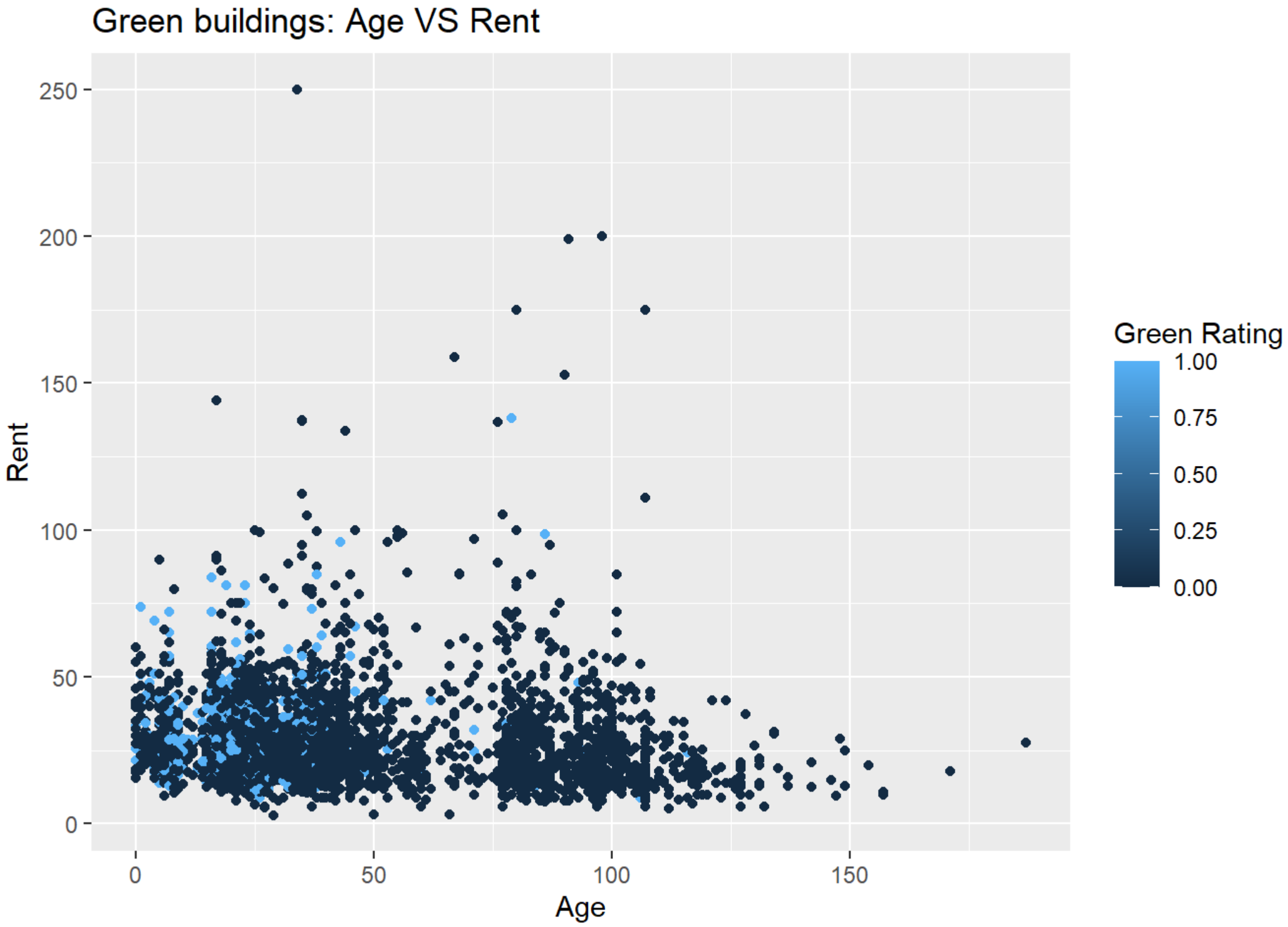
Above, it appears as before. Energystar rated buildings are situated lower than non-energystar rated buildings in terms of rent.

```
## Warning: Removed 74 rows containing missing values (geom_point).
```

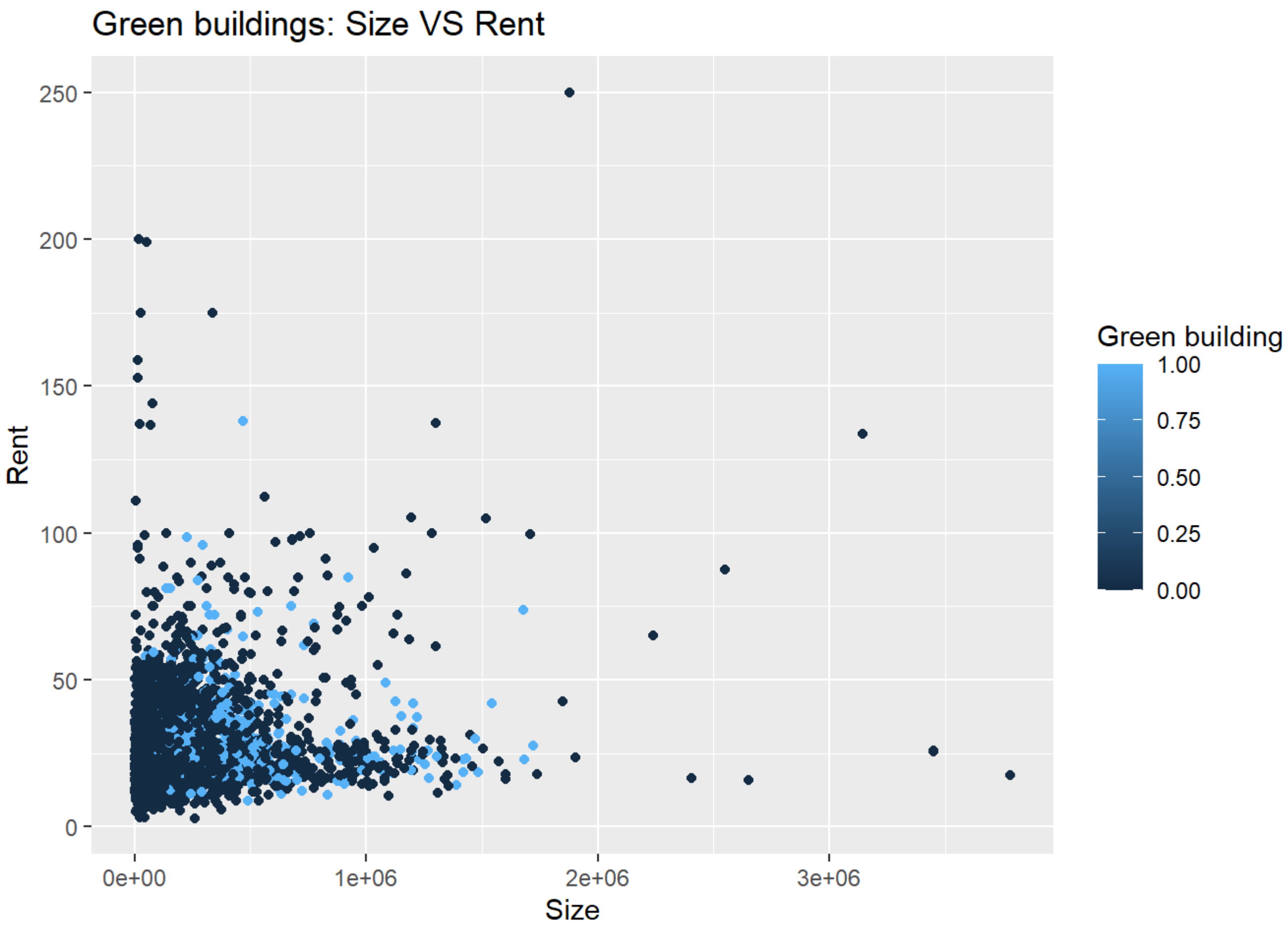


I am

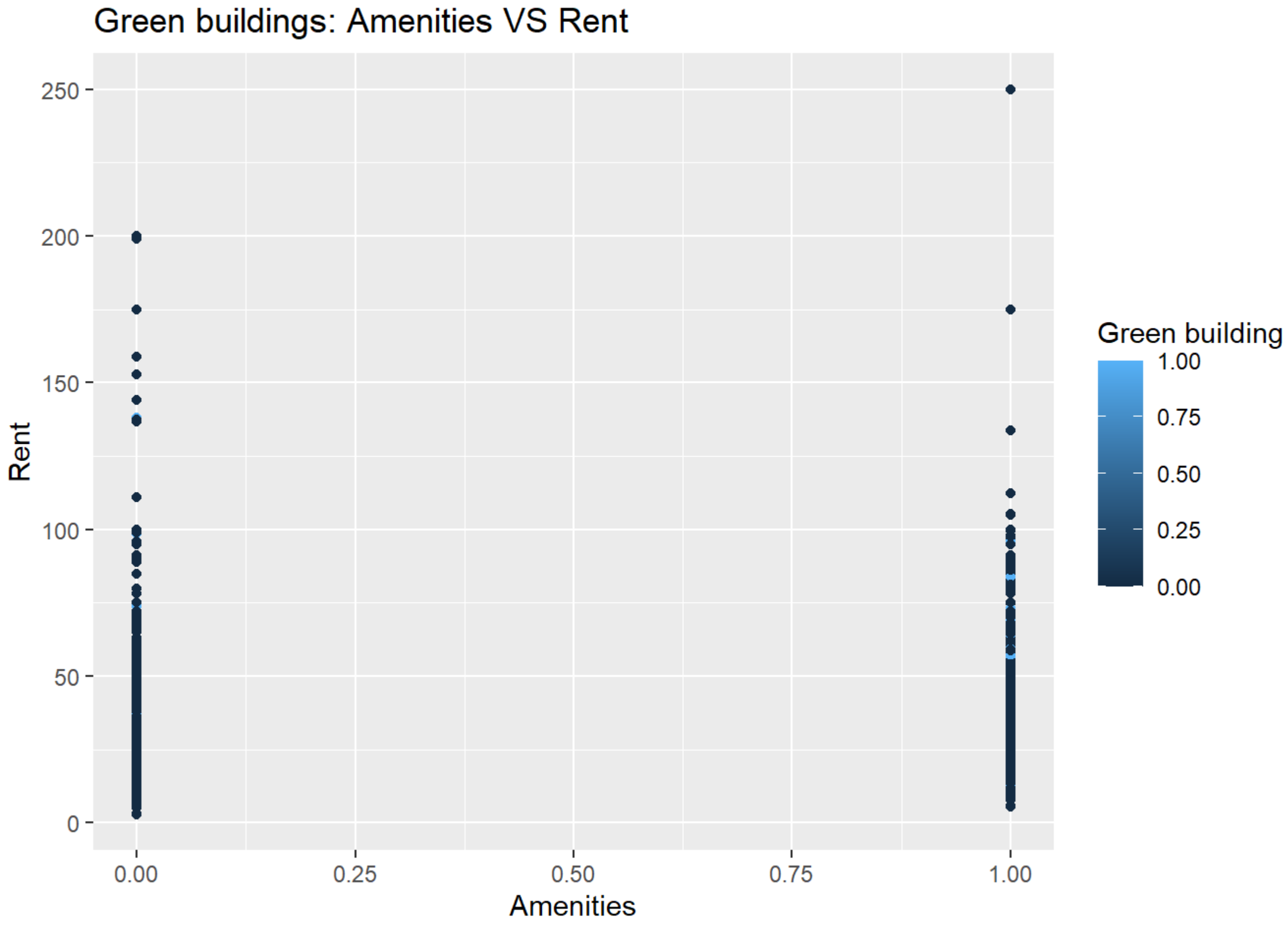
searching for a feature that would increase rent premiums. It does not appear that markets of higher growth influence rent.



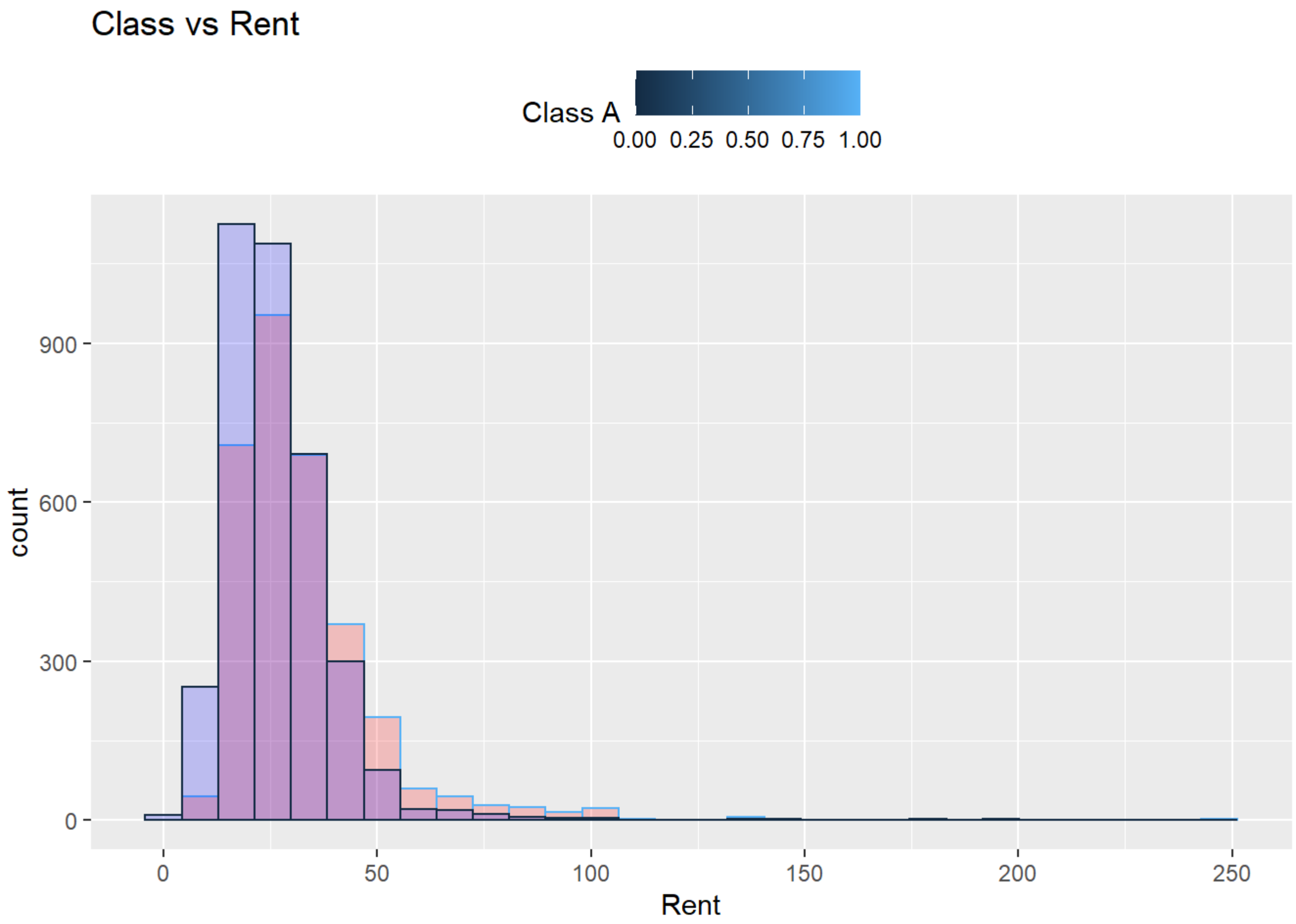
It appears that age does not influence rent all that much. As shown by the light blue dots, most energy efficient buildings are much newer.



It does not seem very conclusive that size of the building is that influential on rent price.

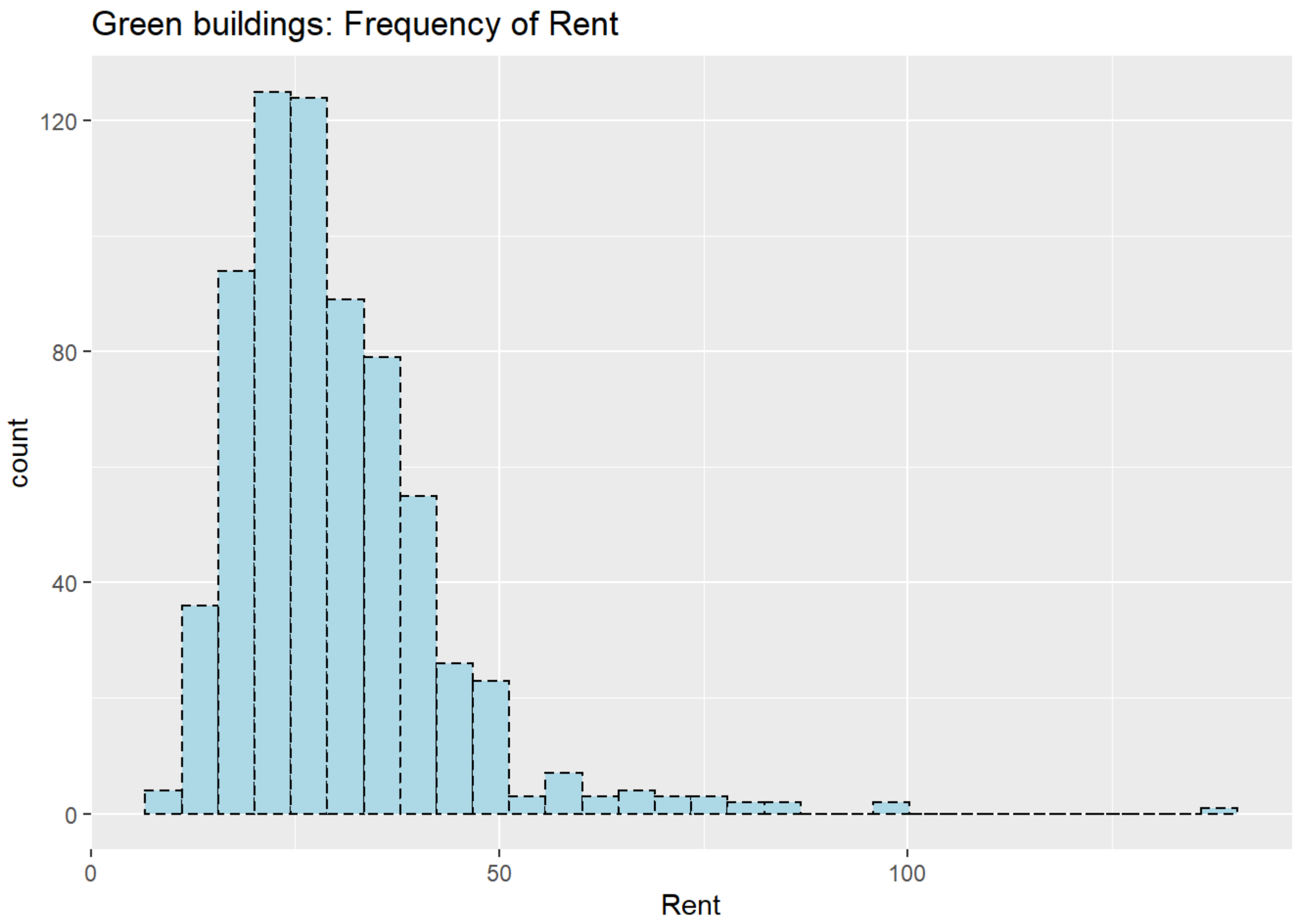


```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



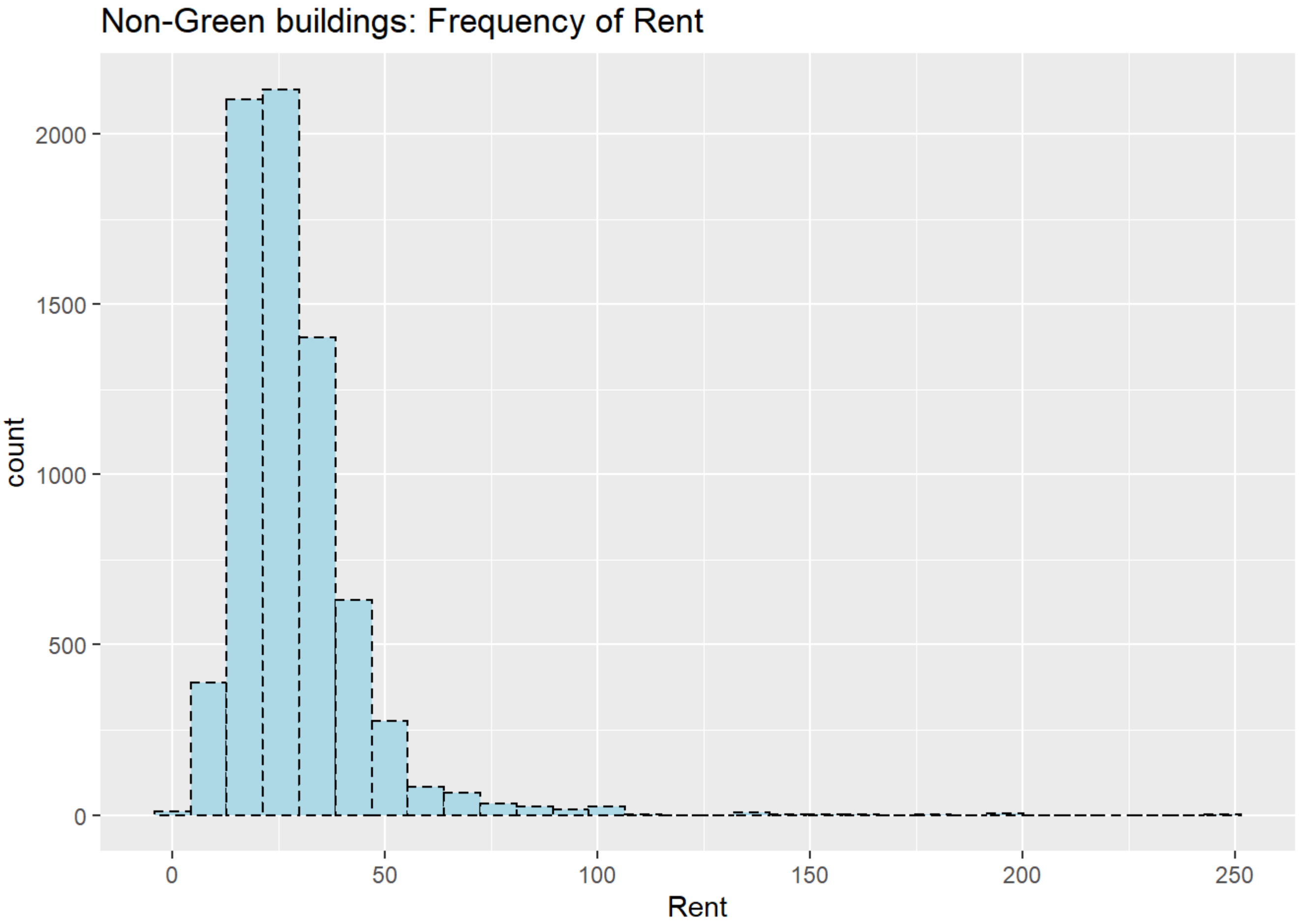
Above, the red bars are class A buildings. The blue are class B buildings. AS you can see, there are premiums for buildings regarded as class A due to the slight skew on the right tail.

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

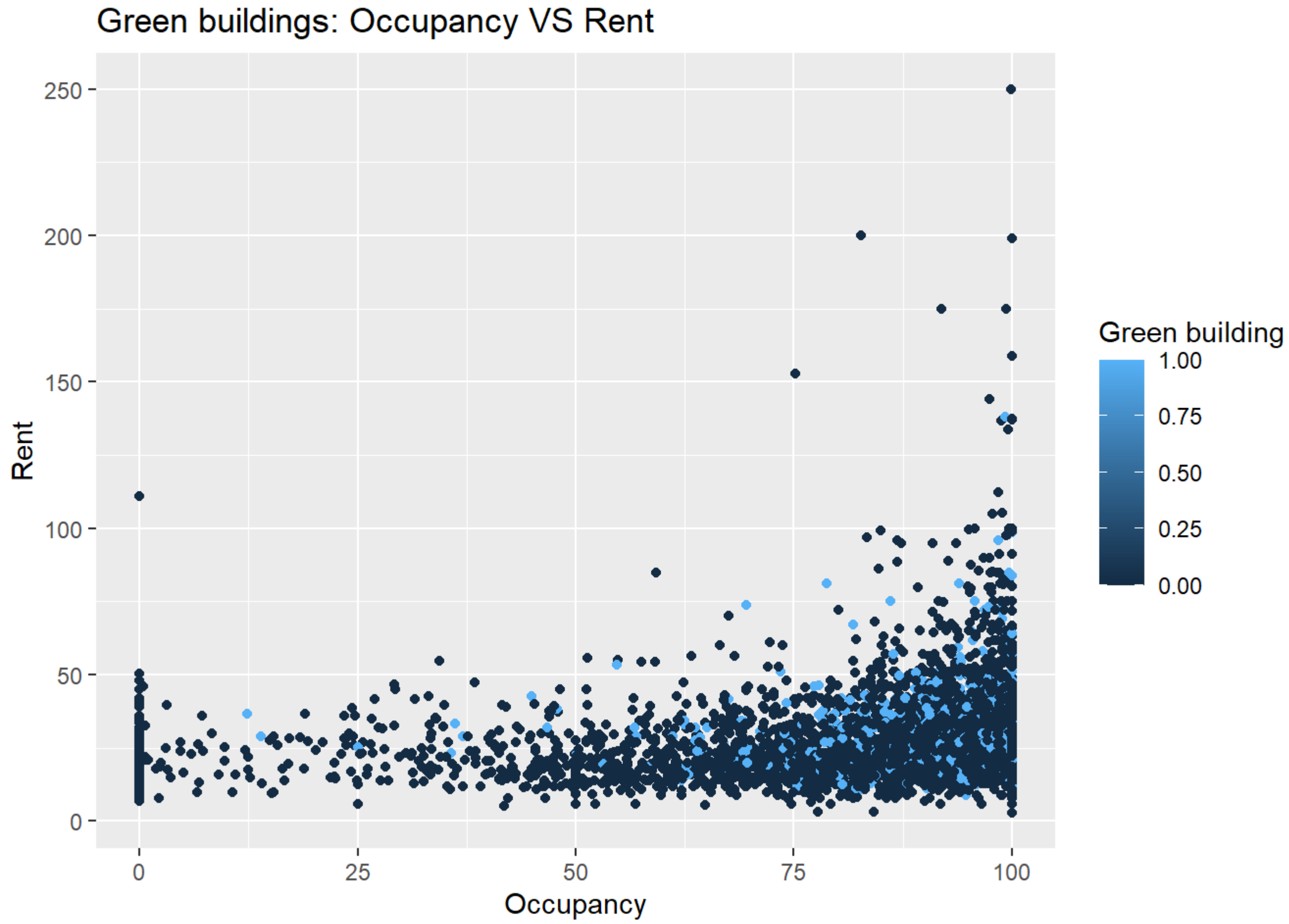


The above chart shows the frequency distribution of rents. This is the primary research the Excel guru used. This individual disregarded the many factors that can influence rents paid. For example, the market that they are in could allow premium rents to be charged. Simply looking at a median value disregards external factors.

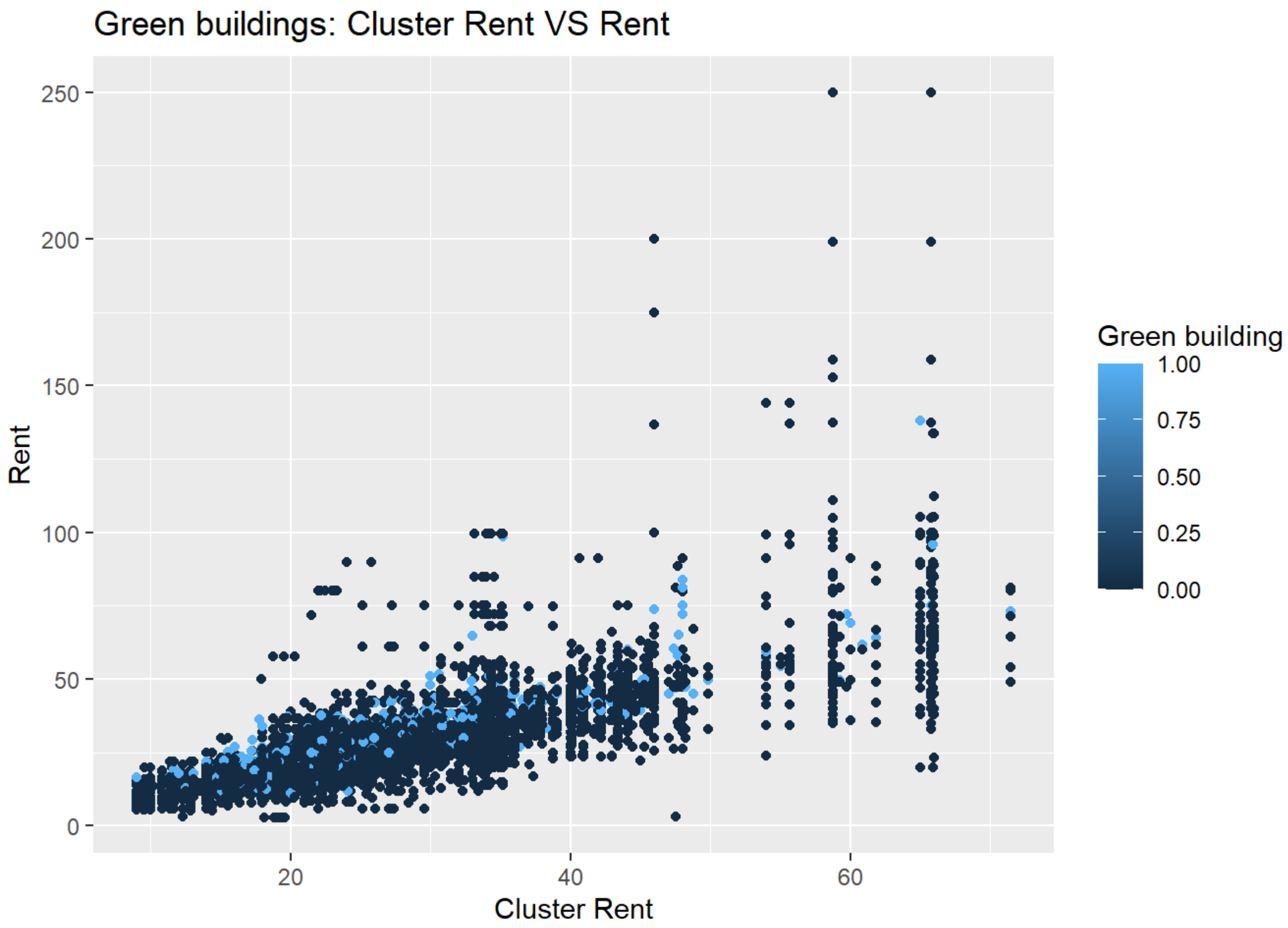
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Above, the Non-Green buildings frequency by rent is showed. There is significant variation. It would be foolish to aggregate and take a median value across these properties to determine what rents could be charged in their area.



Looking at occupancy versus rent, it seems that properties with higher occupancy have higher rent. This is perhaps a confounding variable because a high occupancy could imply there is high demand for a property as such. We should explore those properties with high occupancy. Clearly these properties are in demand, and we should identify these properties as ideals.



It seems that the most significant relationship between rent and a factor is what the average market rent in the area is currently at. This makes sense as you will want to charge the market going rate. Additionally, it does seem that the green buildings, colored in light blue, sit at a premium, on average, to their inefficient counterparts.

So the excel guru really missed out on key factors like understanding what market rents are across different markets. This is the most influential factor that goes into rent expectations. Without running regressions, we cannot be completely sure exactly how the relationship between the variables and rents. Simply grabbing median values as your rent expectation disregards many other factors that influence what a building could charge in rent.

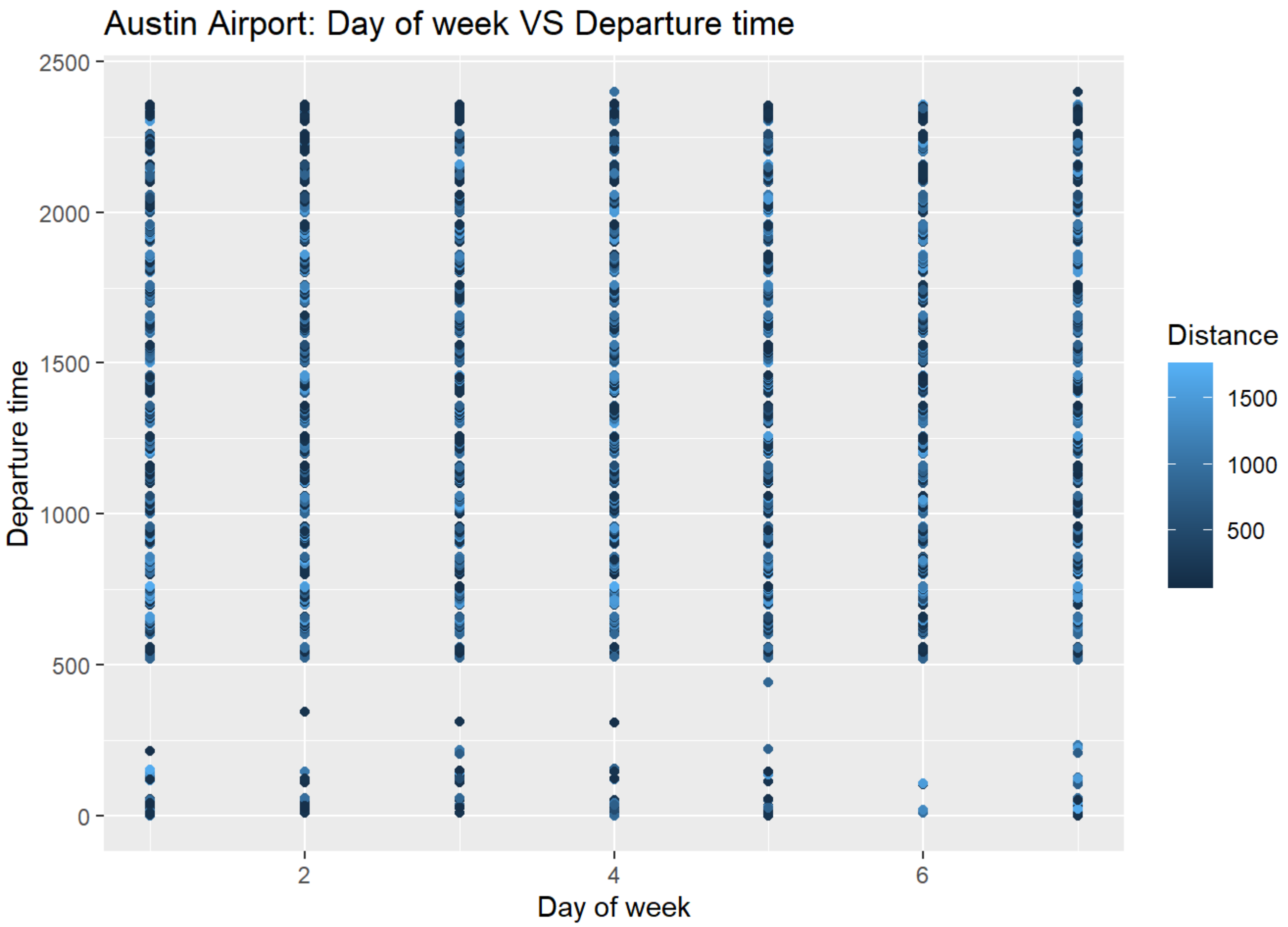
Additionally, the fact that green buildings are typically newer, the comparison between average non-green buildings is a bad comparison. The analyst needs to go into more granular detail and identify properties within her market and similar ages of properties of similar class rating.

#Visual story telling part 2: flights at ABIA

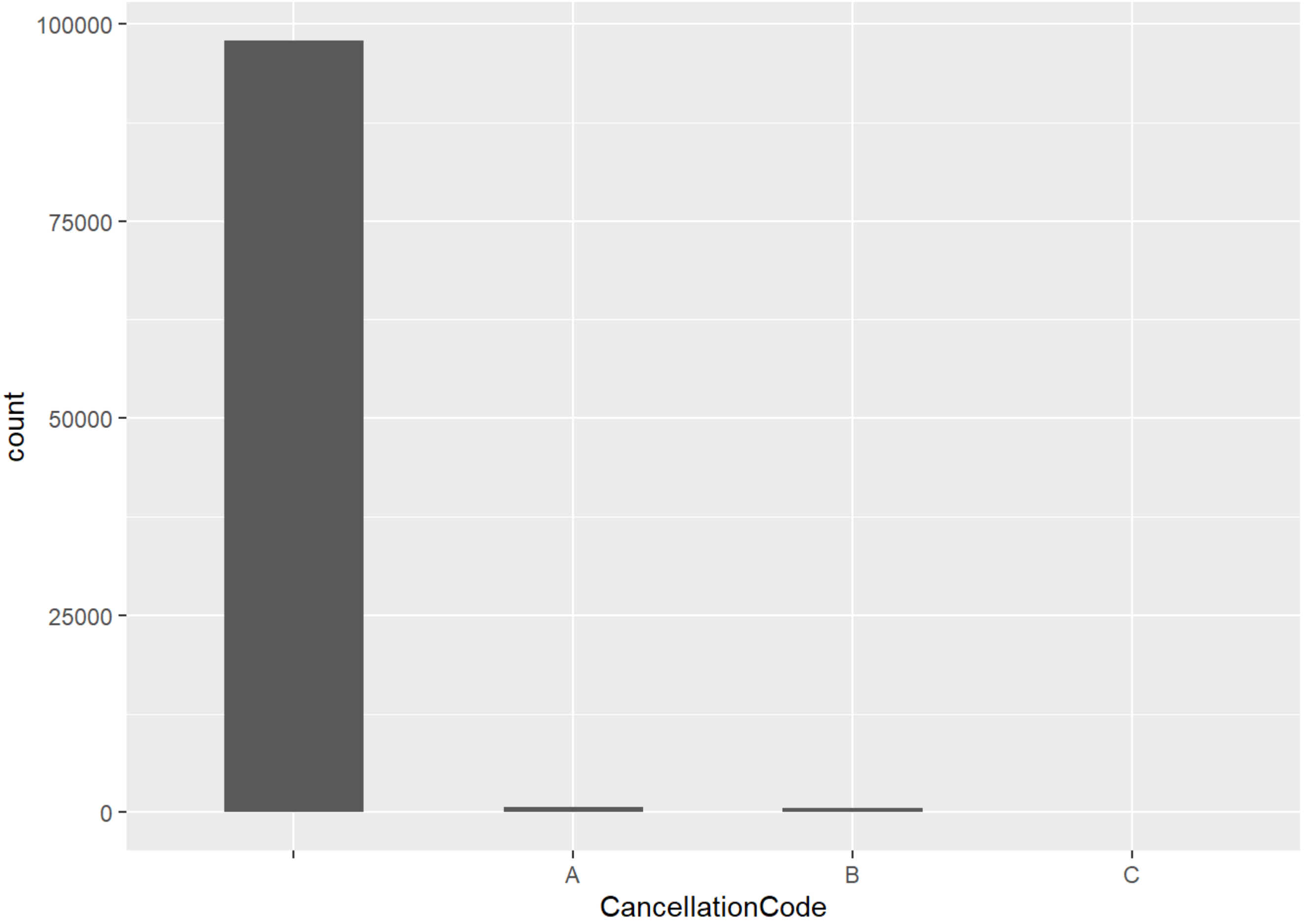
```
##       Year          Month          DayofMonth      DayOfWeek        DepTime    
##  Min.   :2008   Min.   : 1.00   Min.   : 1.00   Min.   :1.000   Min.   :   1  
##  1st Qu.:2008   1st Qu.: 3.00   1st Qu.: 8.00   1st Qu.:2.000   1st Qu.: 917  
##  Median :2008   Median : 6.00   Median :16.00   Median :4.000   Median :1329  
##  Mean   :2008   Mean   : 6.29   Mean   :15.73   Mean   :3.902   Mean   :1329  
##  3rd Qu.:2008   3rd Qu.: 9.00   3rd Qu.:23.00   3rd Qu.:6.000   3rd Qu.:1728  
##  Max.   :2008   Max.   :12.00   Max.   :31.00   Max.   :7.000   Max.   :2400  
##                                                                 NA's   :1413  
##    CRSDepTime      ArrTime       CRSArrTime   UniqueCarrier        FlightNum   
##  Min.   :  55   Min.   :   1   Min.   :   5   Length:99260       Min.   :   1  
##  1st Qu.: 915   1st Qu.:1107   1st Qu.:1115   Class :character   1st Qu.: 640  
##  Median :1320   Median :1531   Median :1535   Mode  :character   Median :1465  
##  Mean   :1320   Mean   :1487   Mean   :1505                      Mean   :1917  
##  3rd Qu.:1720   3rd Qu.:1903   3rd Qu.:1902                      3rd Qu.:2653  
##  Max.   :2346   Max.   :2400   Max.   :2400                      Max.   :9741  
##                 NA's   :1567                                                  
##    TailNum          ActualElapsedTime CRSElapsedTime     AirTime      
##  Length:99260       Min.   : 22.0     Min.   : 17.0   Min.   :  3.00  
##  Class :character   1st Qu.: 57.0     1st Qu.: 58.0   1st Qu.: 38.00  
##  Mode  :character   Median :125.0     Median :130.0   Median :105.00  
##                     Mean   :120.2     Mean   :122.1   Mean   : 99.81  
##                     3rd Qu.:164.0     3rd Qu.:165.0   3rd Qu.:142.00  
##                     Max.   :506.0     Max.   :320.0   Max.   :402.00  
##                     NA's   :1601      NA's   :11      NA's   :1601    
##     ArrDelay           DepDelay          Origin              Dest          
##  Min.   :-129.000   Min.   :-42.000   Length:99260       Length:99260      
##  1st Qu.:  -9.000   1st Qu.: -4.000   Class :character   Class :character  
##  Median :  -2.000   Median :  0.000   Mode  :character   Mode  :character  
##  Mean   :   7.065   Mean   :  9.171                                        
##  3rd Qu.:  10.000   3rd Qu.:  8.000                                        
##  Max.   : 948.000   Max.   :875.000                                        
##  NA's   :1601       NA's   :1413                                           
##     Distance        TaxiIn           TaxiOut         Cancelled      
##  Min.   :  66   Min.   :  0.000   Min.   :  1.00   Min.   :0.00000  
##  1st Qu.: 190   1st Qu.:  4.000   1st Qu.:  9.00   1st Qu.:0.00000  
##  Median : 775   Median :  5.000   Median : 12.00   Median :0.00000  
##  Mean   : 705   Mean   :  6.413   Mean   : 13.96   Mean   :0.01431  
##  3rd Qu.:1085   3rd Qu.:  7.000   3rd Qu.: 16.00   3rd Qu.:0.00000  
##  Max.   :1770   Max.   :143.000   Max.   :305.00   Max.   :1.00000  
##                 NA's   :1567      NA's   :1419                      
##  CancellationCode      Diverted          CarrierDelay     WeatherDelay    
##  Length:99260       Min.   :0.000000   Min.   :  0.00   Min.   :  0.00   
##  Class :character   1st Qu.:0.000000   1st Qu.:  0.00   1st Qu.:  0.00   
##  Mode  :character   Median :0.000000   Median :  0.00   Median :  0.00   
##                     Mean   :0.001824   Mean   : 15.39   Mean   :  2.24   
##                     3rd Qu.:0.000000   3rd Qu.: 16.00   3rd Qu.:  0.00   
##                     Max.   :1.000000   Max.   :875.00   Max.   :412.00   
##                                        NA's   :79513    NA's   :79513    
##     NASDelay      SecurityDelay    LateAircraftDelay
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   
##  1st Qu.:  0.00   1st Qu.:  0.00   1st Qu.:  0.00   
##  Median :  2.00   Median :  0.00   Median :  6.00   
##  Mean   : 12.47   Mean   :  0.07   Mean   : 22.97   
##  3rd Qu.: 16.00   3rd Qu.:  0.00   3rd Qu.: 30.00  
##  Max.   :367.00   Max.   :199.00   Max.   :458.00  
##  NA's   :79513    NA's   :79513    NA's   :79513
```

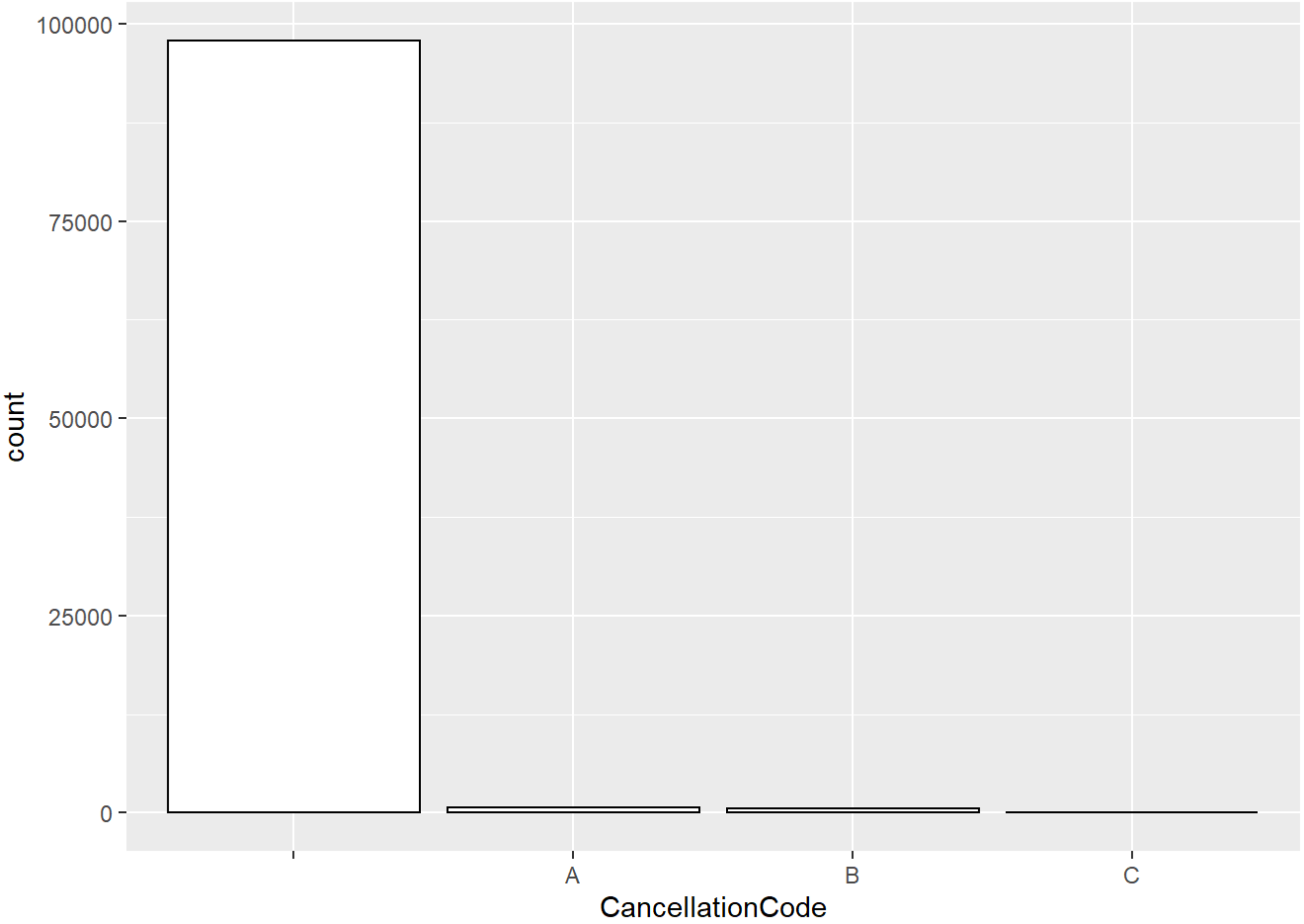
The above output are summary statistics for all features of the ABIA data set. There is a mix between categorical variables and quantitative variables.

```
## Warning: Removed 1413 rows containing missing values (geom_point).
```

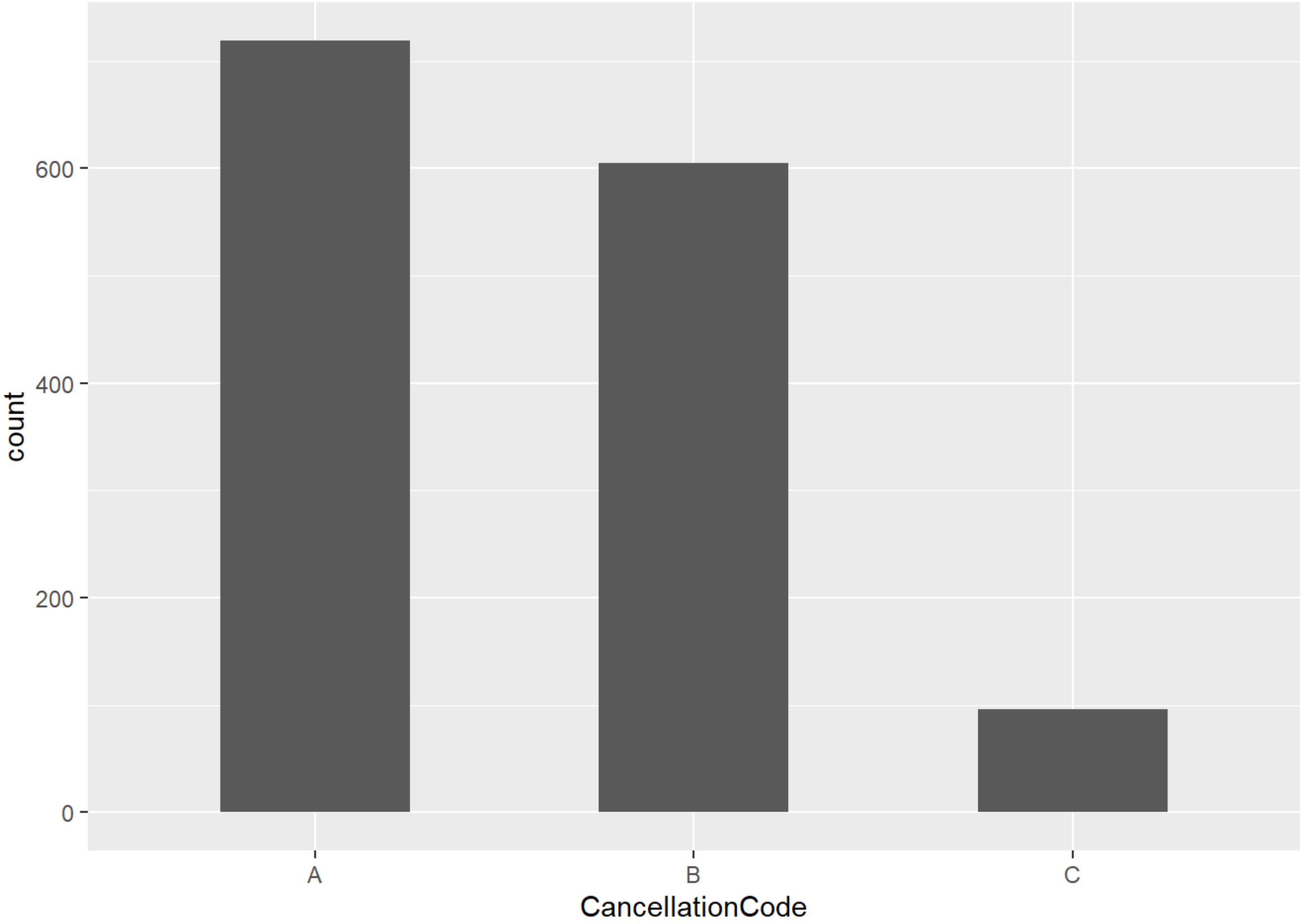


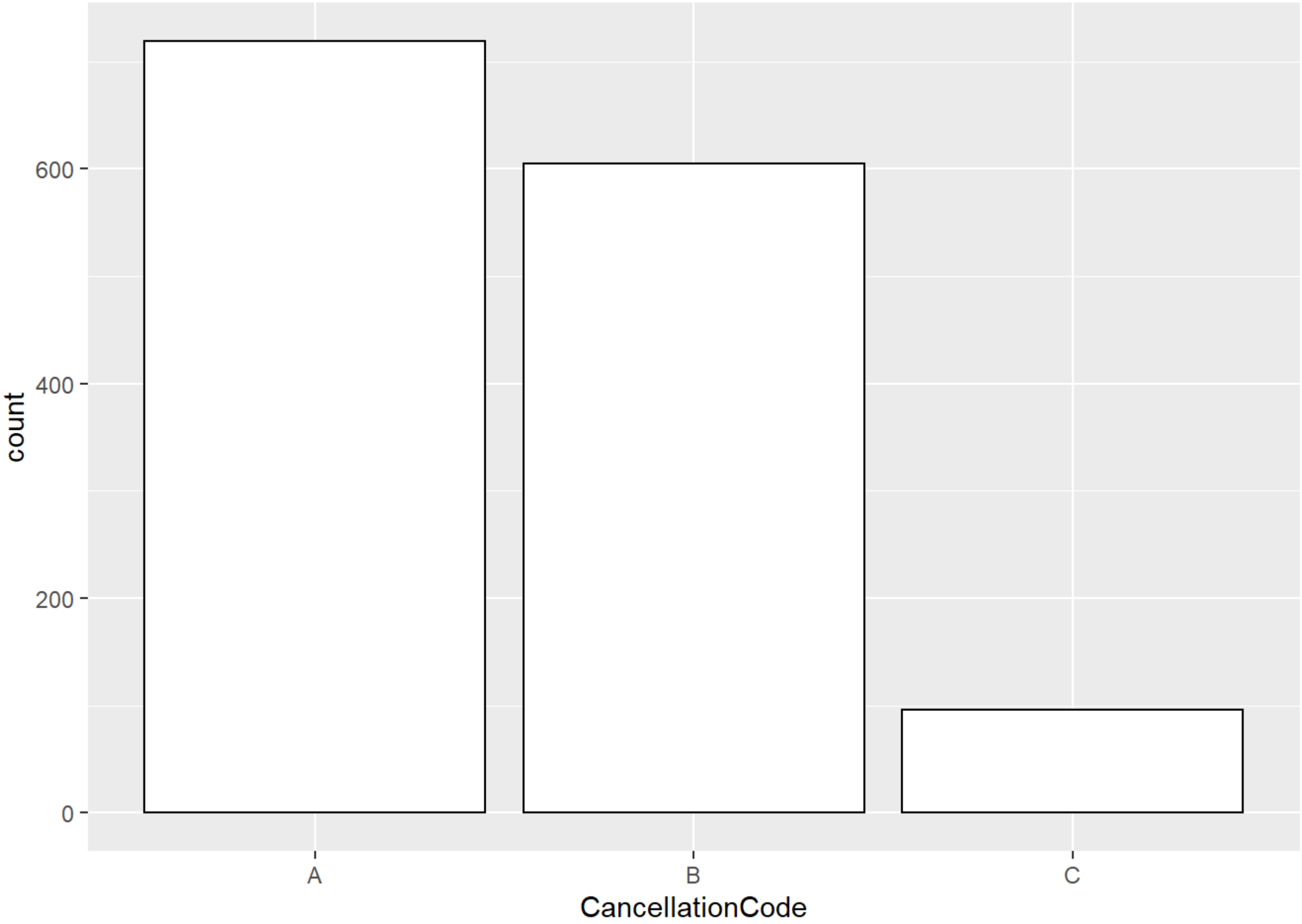
Looking at the chart above, the Austin Airport has many flights that occur daily. With the color mapping to the distance of the flight, it does not appear that there is any apparent relationship between the day, time, and distance of the flight.





Most planes do not have cancellation codes.

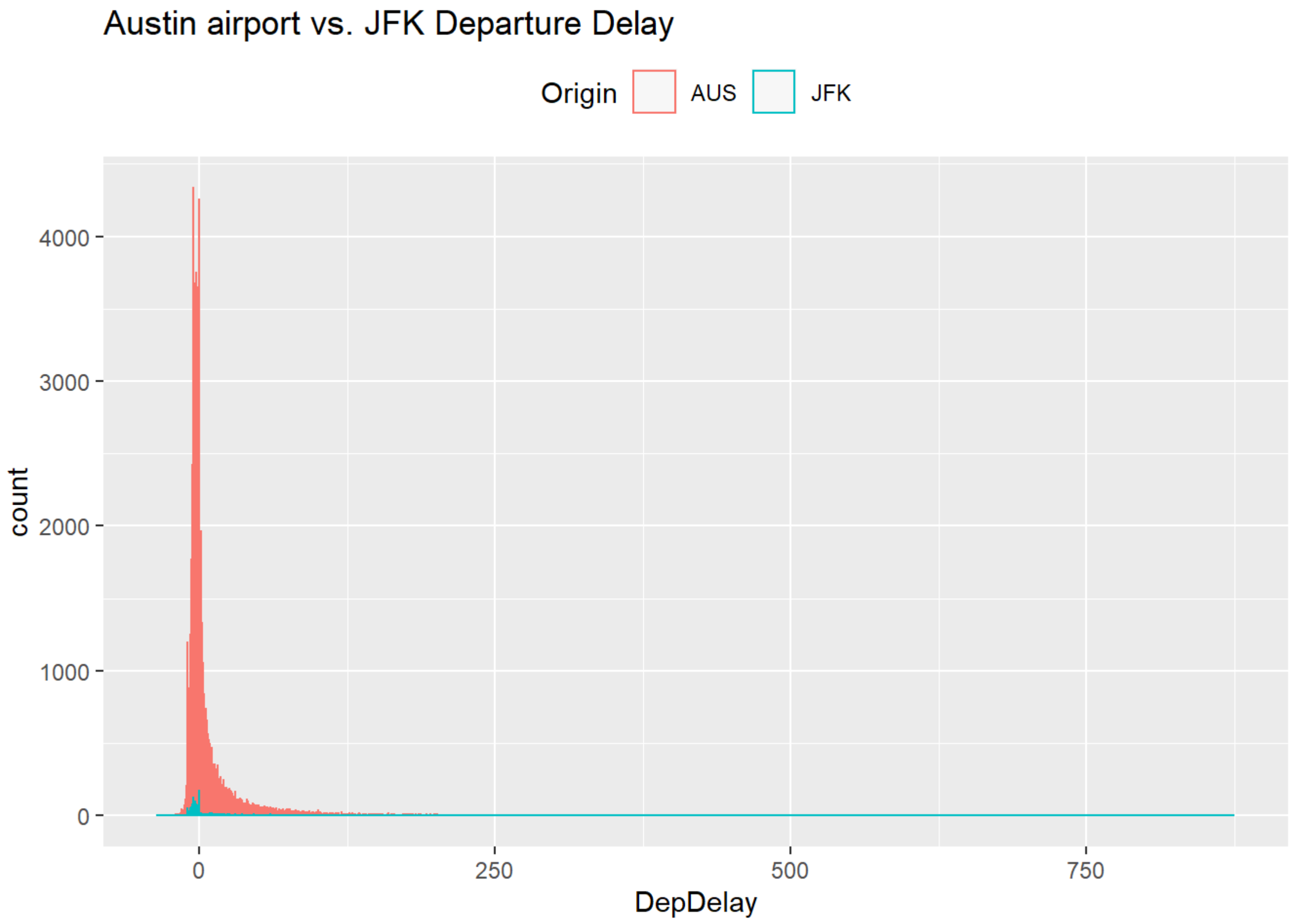




Of the flights that were canceled, the primary reason was due to carrier issues. In close second, flights are canceled due to weather. In third, flights are canceled due to NAS. There were no instances of “D” security cancellations.

```
## Warning: Removed 730 rows containing non-finite values (stat_bin).
```

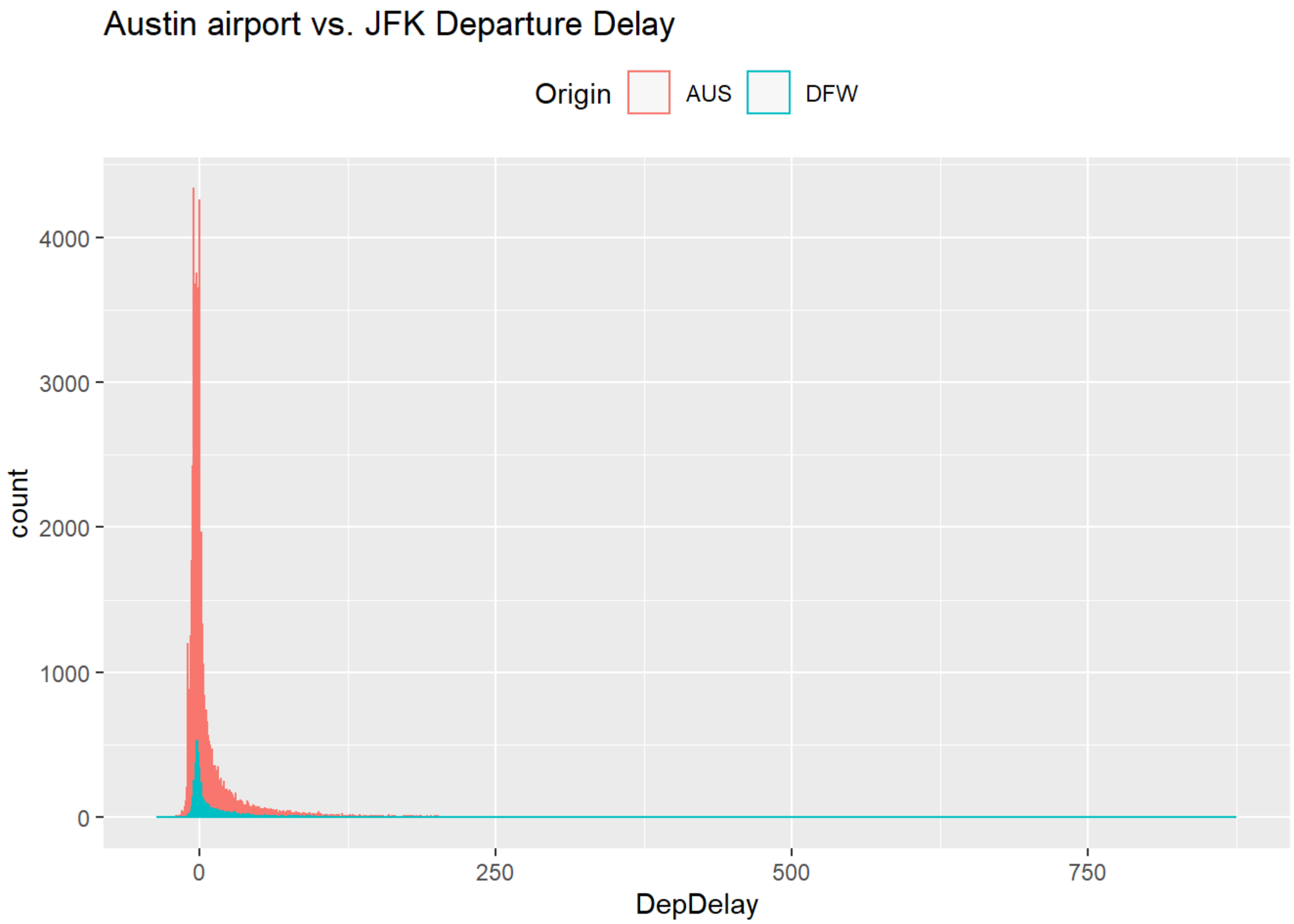
```
## Warning: Removed 33 rows containing non-finite values (stat_bin).
```



Compared to JFK, the Austin airport has significantly more departure delays than that of JFK. Expect delays if you are going anywhere from Austin. This could likely be due to the weather difference in Austin to New York.

```
## Warning: Removed 730 rows containing non-finite values (stat_bin).
```

```
## Warning: Removed 158 rows containing non-finite values (stat_bin).
```



Interestingly, the Dallas-Fort Worth airport has a similar distribution to Austin’s. Perhaps this is due to shared weather delays.

#Portfolio modeling

```
## 'getSymbols' currently uses auto.assign=TRUE by default, but will
## use auto.assign=FALSE in 0.5-0. You will still be able to use
## 'loadSymbols' to automatically load data. getOption("getSymbols.env")
## and getOption("getSymbols.auto.assign") will still be checked for
## alternate defaults.
##
## This message is shown once per session and may be disabled by setting
## options("getSymbols.warning4.0"=FALSE). See ?getSymbols for details.
```

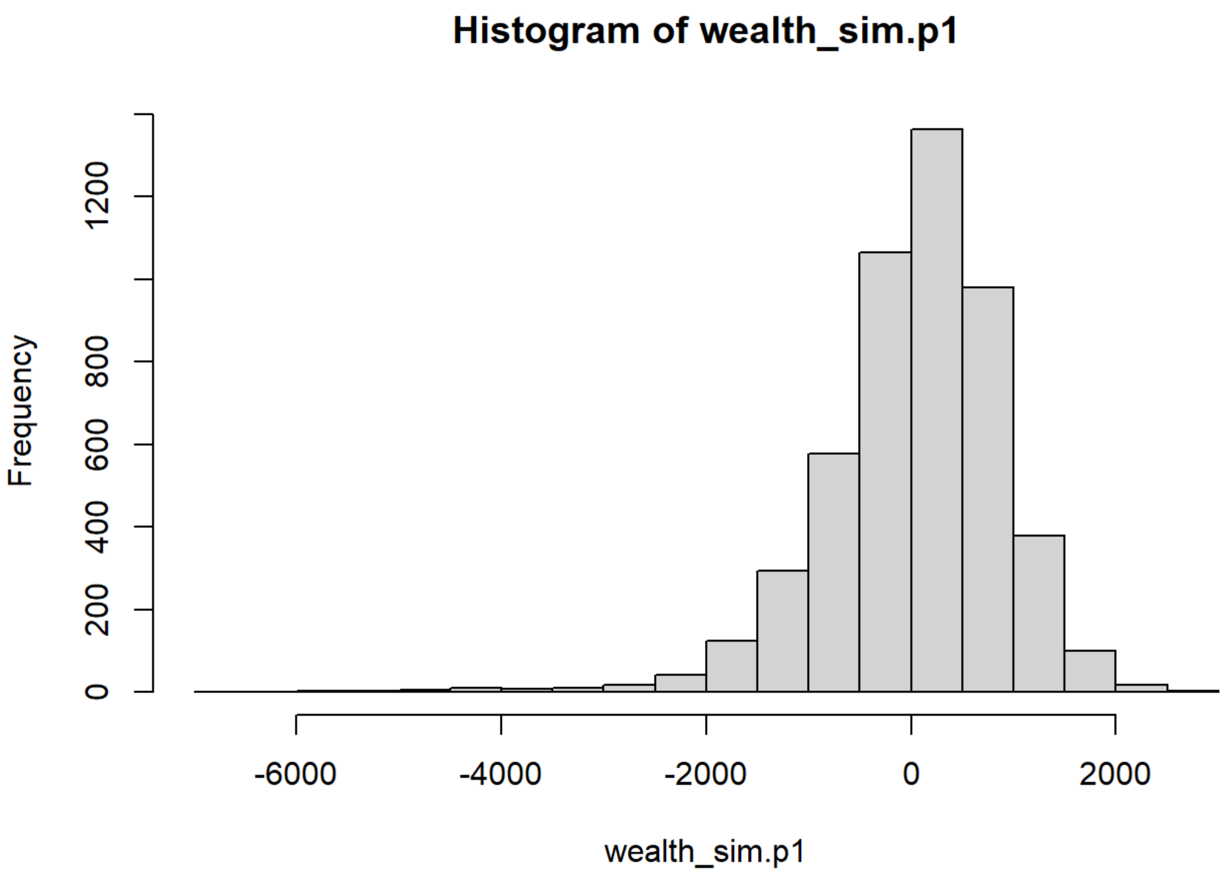
```
## [1] "VGK"
```

```
## [1] "EWU"
```

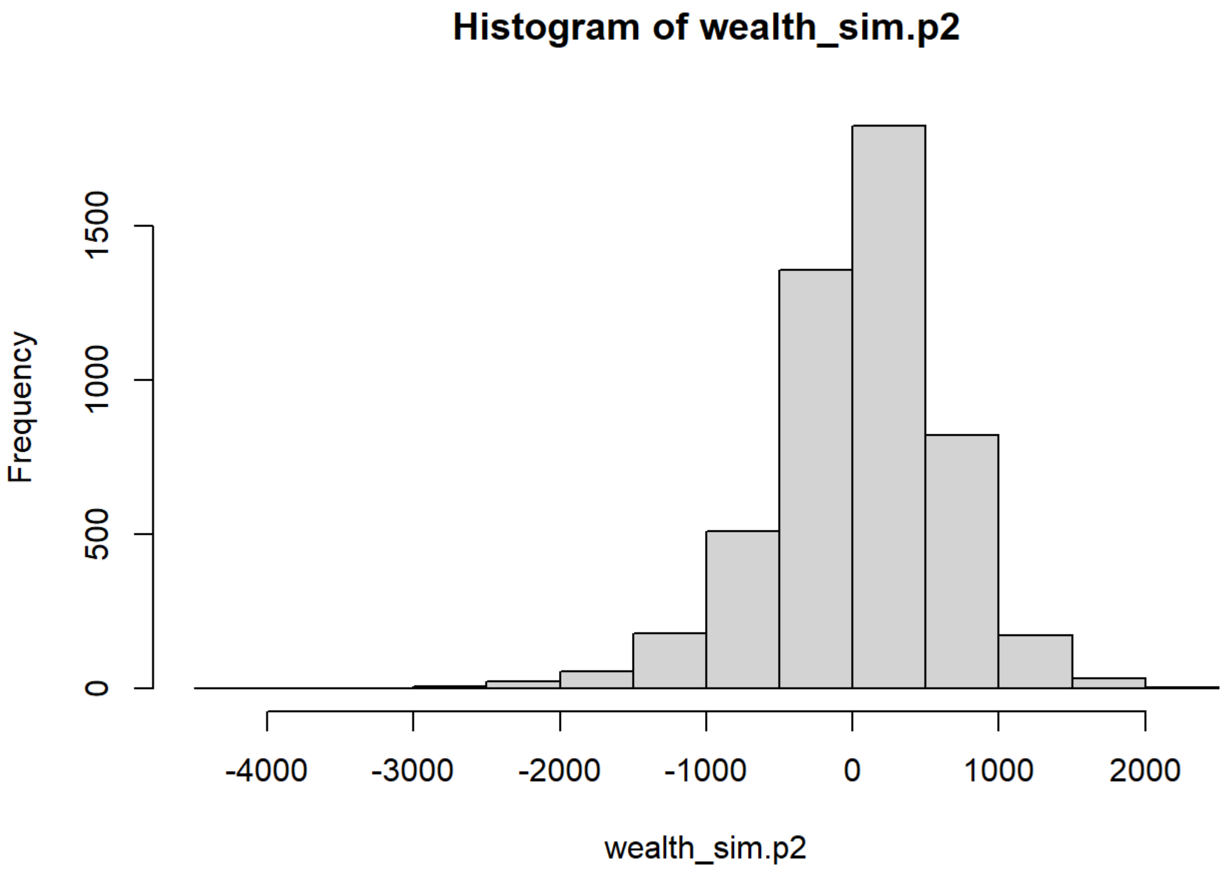
```
## [1] "SHYG"
```

```
## [1] "URE"
```

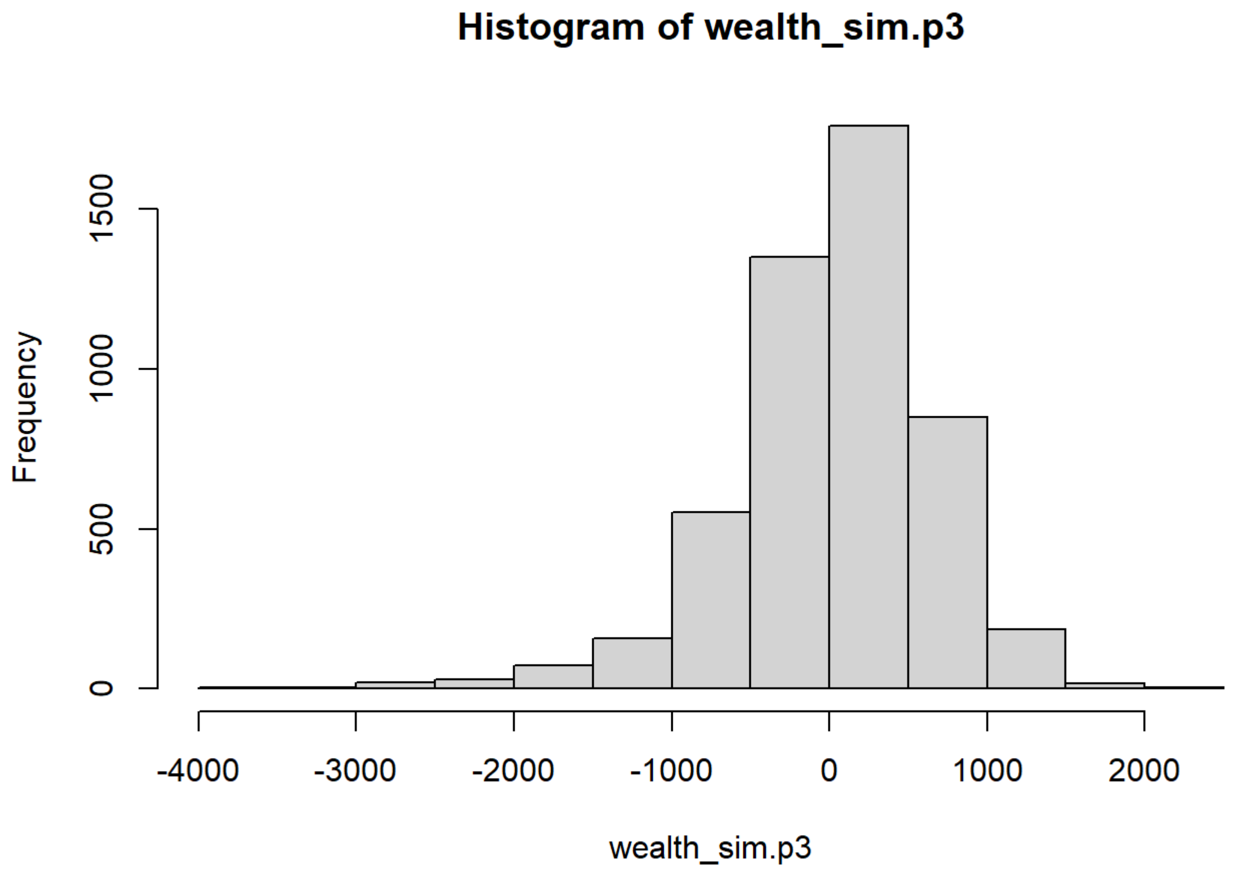
```
## [1] "SVXY"
##        5%
## -1431.915
```



```
##        5%
## -1029.164
```


Histogram of wealth_sim.p2
Frequency
0
500
1000
1500
-4000
-3000
-2000
-1000
0
1000
2000
wealth_sim.p2

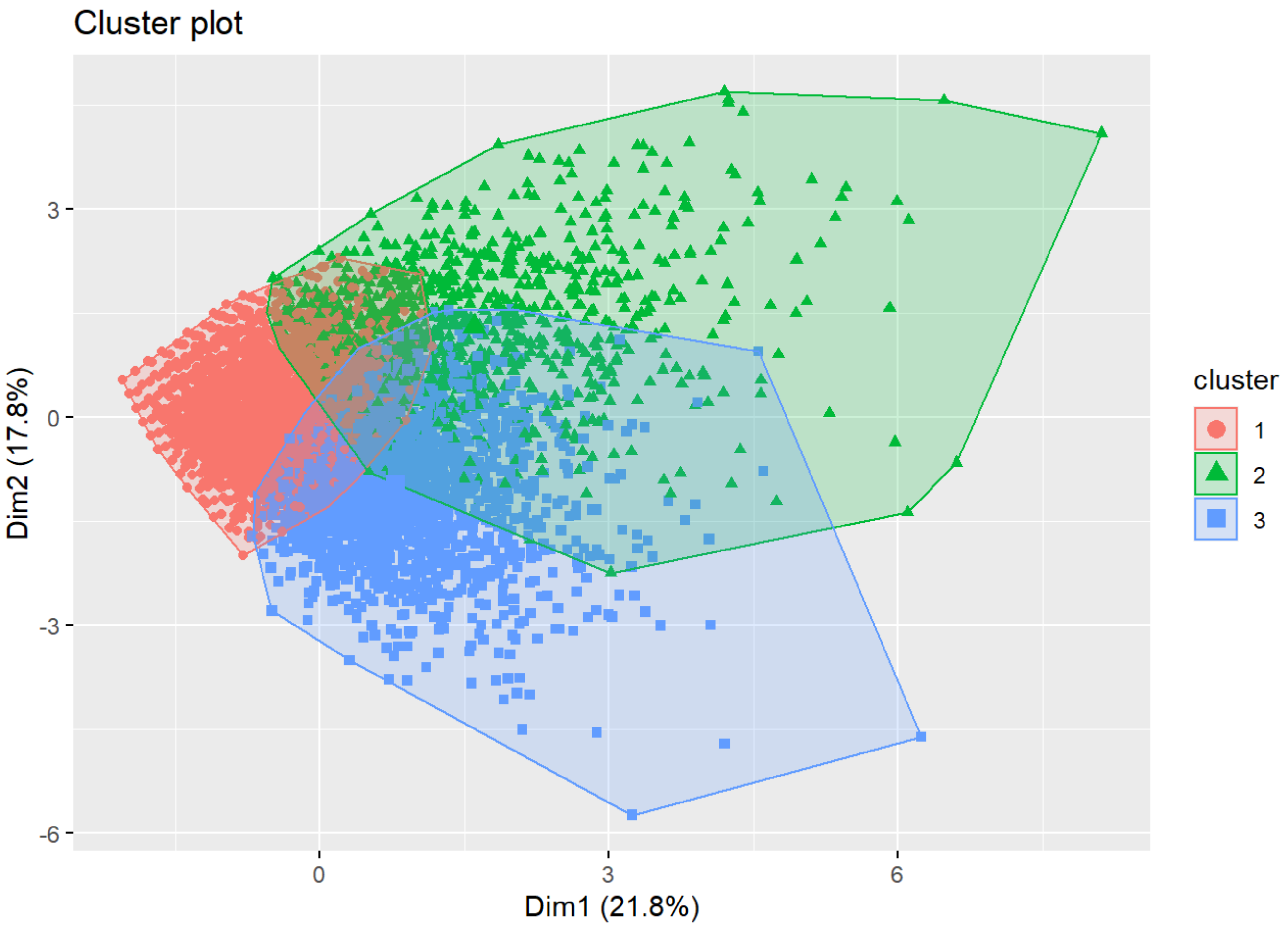
```
##       5%
## -1078.14
```



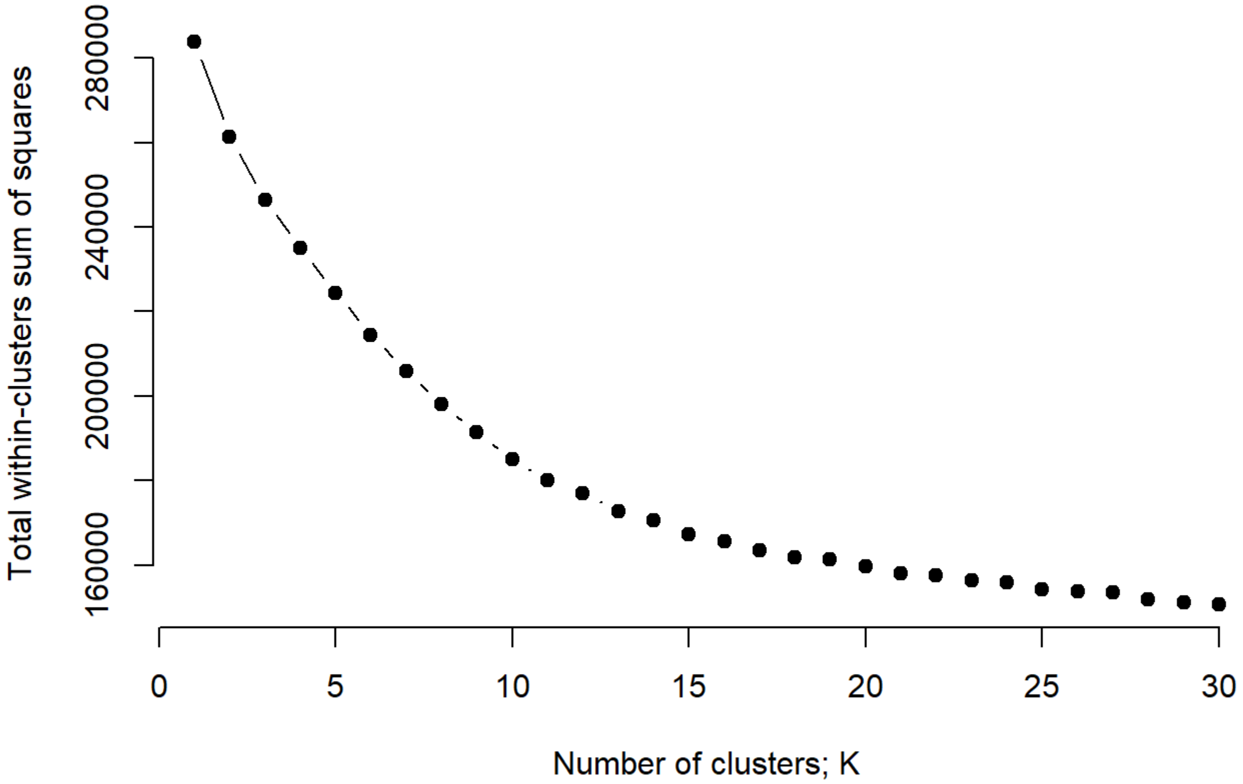
Some of these portfolio returns are very skewed! Asset returns appear to be non-normal in many cases.

#Market segmentation

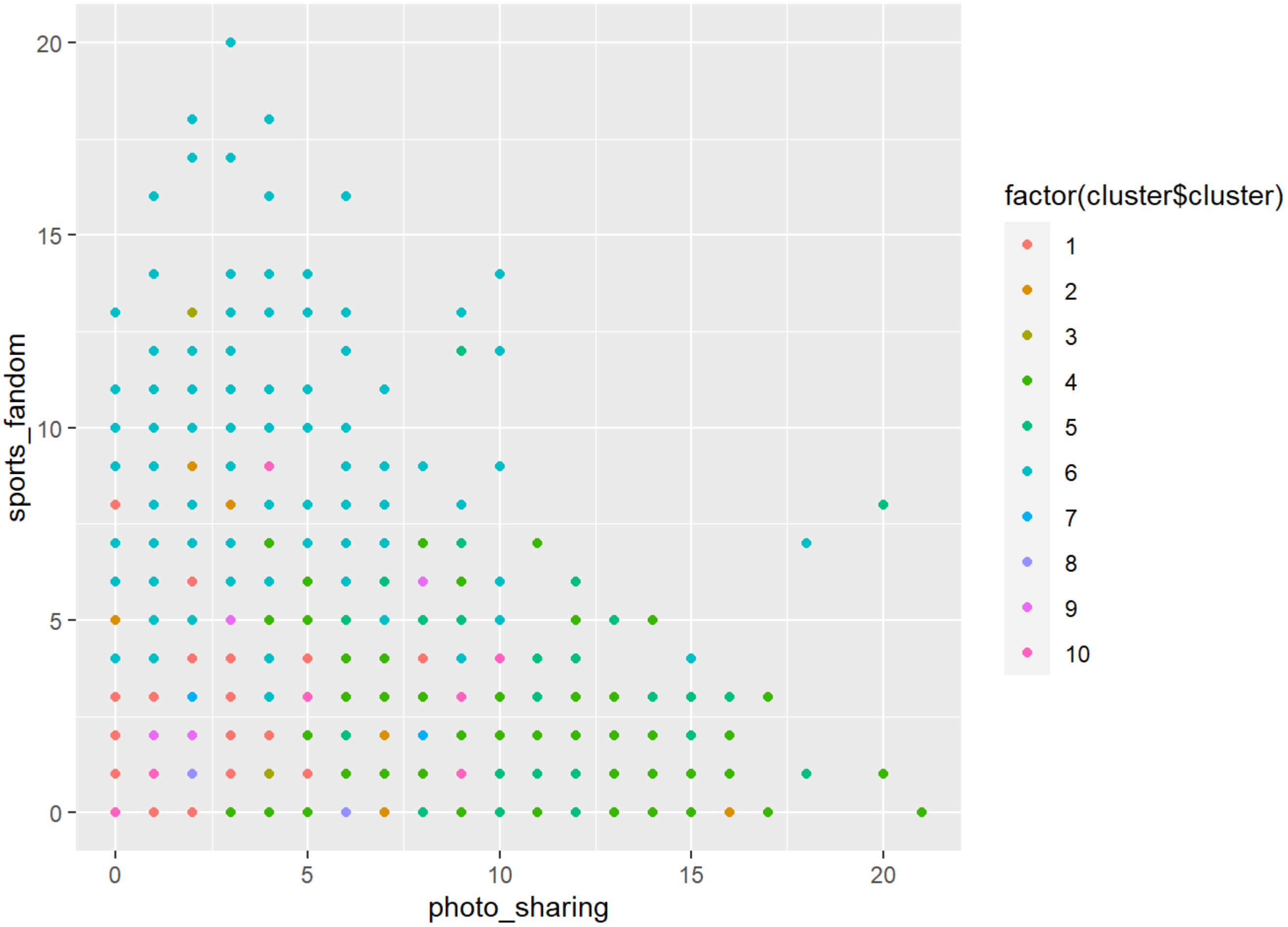
```
##       X               chatter       current_events      travel
##  Length:7882        Min.   : 0.000   Min.   :0.000   Min.   : 0.000
##  Class :character   1st Qu.: 2.000   1st Qu.:1.000   1st Qu.: 0.000
##  Mode  :character   Median : 3.000   Median :1.000   Median : 1.000
##                     Mean   : 4.399   Mean   :1.526   Mean   : 1.585
##                     3rd Qu.: 6.000   3rd Qu.:2.000   3rd Qu.: 2.000
##                     Max.   :26.000   Max.   :8.000   Max.   :26.000
##  photo_sharing    uncategorized      tv_film      sports_fandom
##  Min.   : 0.000   Min.   :0.000   Min.   : 0.00   Min.   : 0.000
##  1st Qu.: 1.000   1st Qu.:0.000   1st Qu.: 0.00   1st Qu.: 0.000
##  Median : 2.000   Median :1.000   Median : 1.00   Median : 1.000
##  Mean   : 2.697   Mean   :0.813   Mean   : 1.07   Mean   : 1.594
##  3rd Qu.: 4.000   3rd Qu.:1.000   3rd Qu.: 1.00   3rd Qu.: 2.000
##  Max.   :21.000   Max.   :9.000   Max.   :17.00   Max.   :20.000
##    politics          food            family        home_and_garden
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:0.0000
##  Median : 1.000   Median : 1.000   Median : 1.0000   Median :0.0000
##  Mean   : 1.789   Mean   : 1.397   Mean   : 0.8639   Mean   :0.5207
##  3rd Qu.: 2.000   3rd Qu.: 2.000   3rd Qu.: 1.0000   3rd Qu.:1.0000
##  Max.   :37.000   Max.   :16.000   Max.   :10.0000   Max.   :5.0000
##      music              news        online_gaming       shopping
##  Min.   : 0.0000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.000
##  1st Qu.: 0.0000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.: 0.000
##  Median : 0.0000   Median : 0.000   Median : 0.000   Median : 1.000
##  Mean   : 0.6793   Mean   : 1.206   Mean   : 1.209   Mean   : 1.389
##  3rd Qu.: 1.0000   3rd Qu.: 1.000   3rd Qu.: 1.000   3rd Qu.: 2.000
##  Max.   :13.0000   Max.   :20.000   Max.   :27.000   Max.   :12.000
##  health_nutrition  college_uni     sports_playing      cooking
##  Min.   : 0.000   Min.   : 0.000   Min.   :0.0000   Min.   : 0.000
##  1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.:0.0000   1st Qu.: 0.000
##  Median : 1.000   Median : 1.000   Median :0.0000   Median : 1.000
##  Mean   : 2.567   Mean   : 1.549   Mean   :0.6392   Mean   : 1.998
##  3rd Qu.: 3.000   3rd Qu.: 2.000   3rd Qu.:1.0000   3rd Qu.: 2.000
##  Max.   :41.000   Max.   :30.000   Max.   :8.0000   Max.   :33.000
##       eco            computers          business         outdoors
##  Min.   :0.0000   Min.   : 0.0000   Min.   :0.0000   Min.   : 0.0000
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.: 0.0000
##  Median :0.0000   Median : 0.0000   Median :0.0000   Median : 0.0000
##  Mean   :0.5123   Mean   : 0.6491   Mean   :0.4232   Mean   : 0.7827
##  3rd Qu.:1.0000   3rd Qu.: 1.0000   3rd Qu.:1.0000   3rd Qu.: 1.0000
##  Max.   :6.0000   Max.   :16.0000   Max.   :6.0000   Max.   :12.0000
##      crafts         automotive          art             religion
##  Min.   :0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.000
##  1st Qu.:0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.000
##  Median :0.0000   Median : 0.0000   Median : 0.0000   Median : 0.000
##  Mean   :0.5159   Mean   : 0.8299   Mean   : 0.7248   Mean   : 1.095
##  3rd Qu.:1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.000
##  Max.   :7.0000   Max.   :13.0000   Max.   :18.0000   Max.   :20.000
##      beauty          parenting          dating            school
##  Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000   Min.   : 0.0000
##  1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000   1st Qu.: 0.0000
##  Median : 0.0000   Median : 0.0000   Median : 0.0000   Median : 0.0000
##  Mean   : 0.7052   Mean   : 0.9213   Mean   : 0.7109   Mean   : 0.7677  
##  3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000   3rd Qu.: 1.0000  
##  Max.   :14.0000   Max.   :14.0000   Max.   :24.0000   Max.   :11.0000  
##  personal_fitness     fashion         small_business        spam        
##  Min.   : 0.000   Min.   : 0.0000   Min.   :0.0000   Min.   :0.00000  
##  1st Qu.: 0.000   1st Qu.: 0.0000   1st Qu.:0.0000   1st Qu.:0.00000  
##  Median : 0.000   Median : 0.0000   Median :0.0000   Median :0.00000  
##  Mean   : 1.462   Mean   : 0.9966   Mean   :0.3363   Mean   :0.00647  
##  3rd Qu.: 2.000   3rd Qu.: 1.0000   3rd Qu.:1.0000   3rd Qu.:0.00000  
##  Max.   :19.000   Max.   :18.0000   Max.   :6.0000   Max.   :2.00000  
##      adult        
##  Min.   : 0.0000  
##  1st Qu.: 0.0000  
##  Median : 0.0000  
##  Mean   : 0.4033  
##  3rd Qu.: 0.0000  
##  Max.   :26.0000  
```



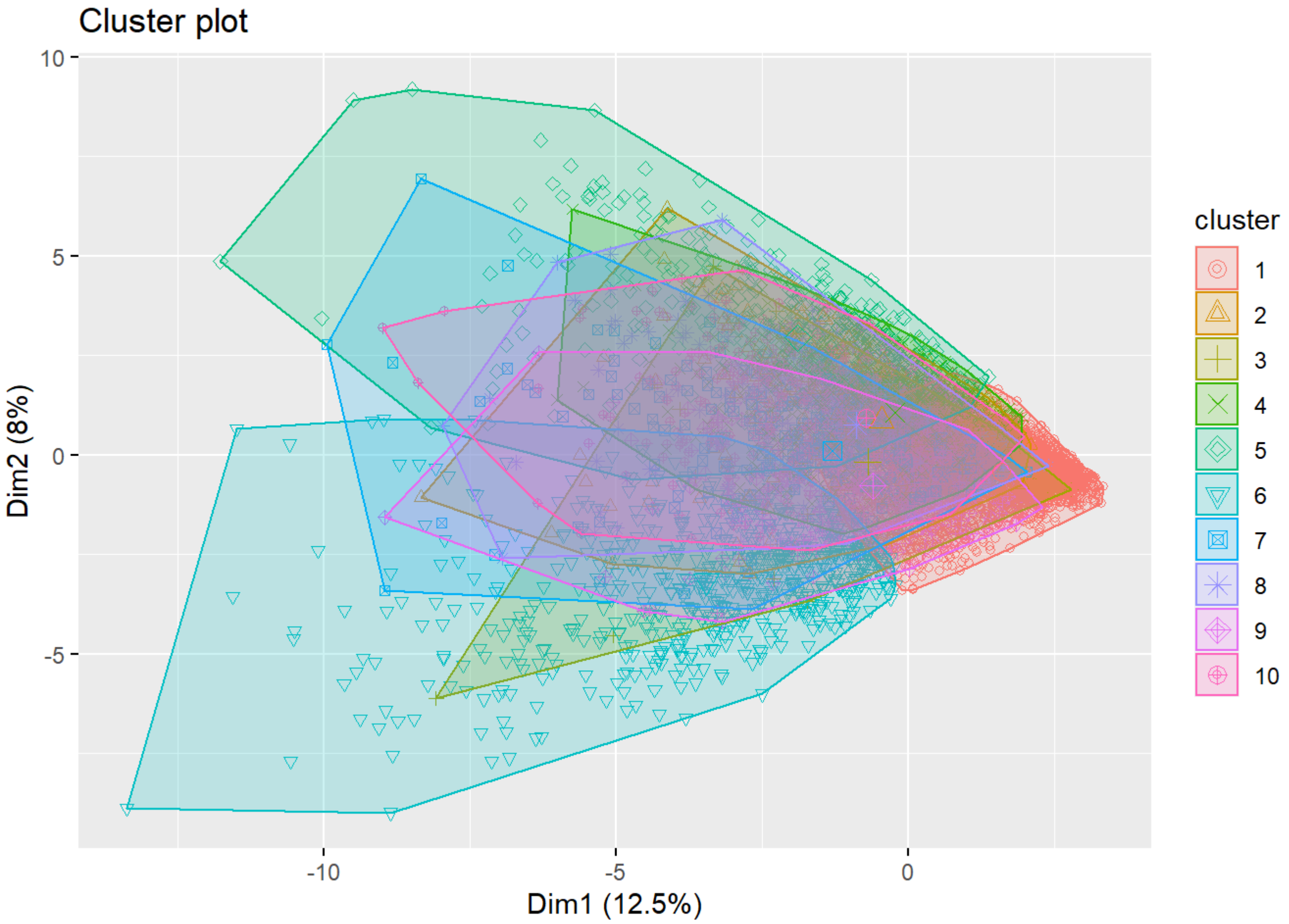
After broadly applying a k-means clustering algorithm to all the features, we achieve the previously shown chart. There is a lot of overlap with three clusters. I attempted to determine what is the optimal number of clusters.



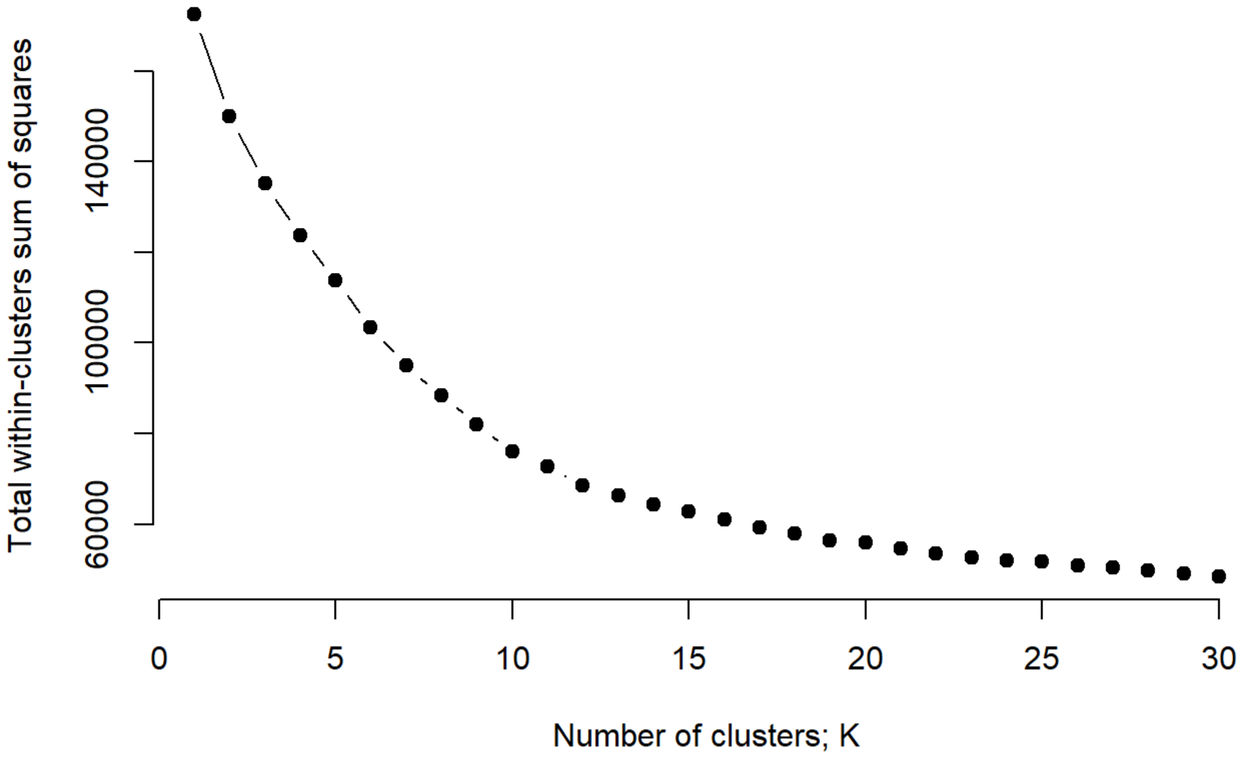
The chart above shows the total within-clusters sum of squares for given K values. K is the number of clusters to used to group the data. As the total within-clusters sum of squares values begin to level off around 20 total clusters, I will proceed with this quantity of clusters.



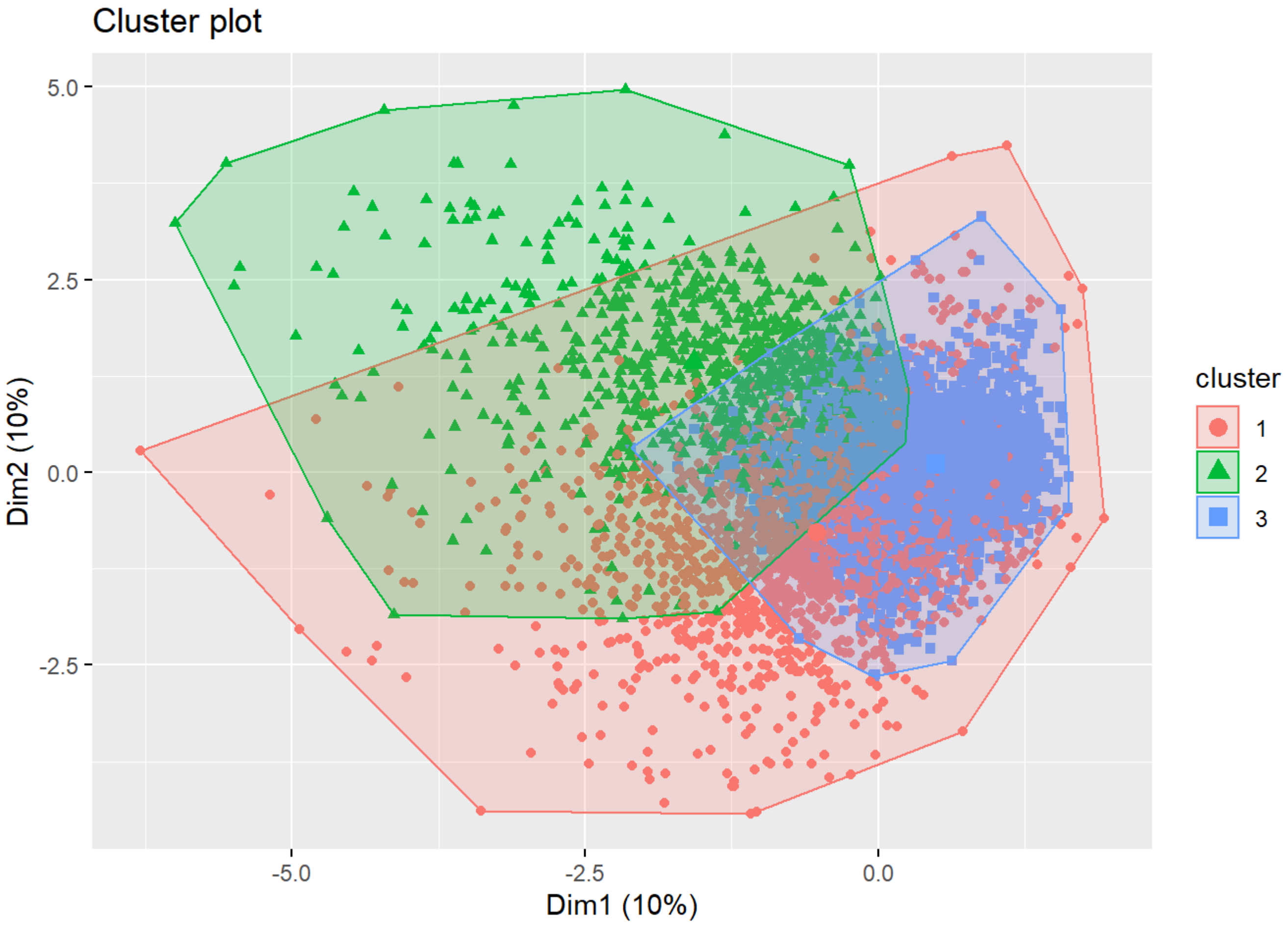
The chart above shows some of the Twitter accounts and their respective cluster classifications. The clustering with 10 clusters does not seem too great. However, we can still see that there are two main clusters I like to think that this clustering model identified subsets of people that enjoy sports or sharing photos.



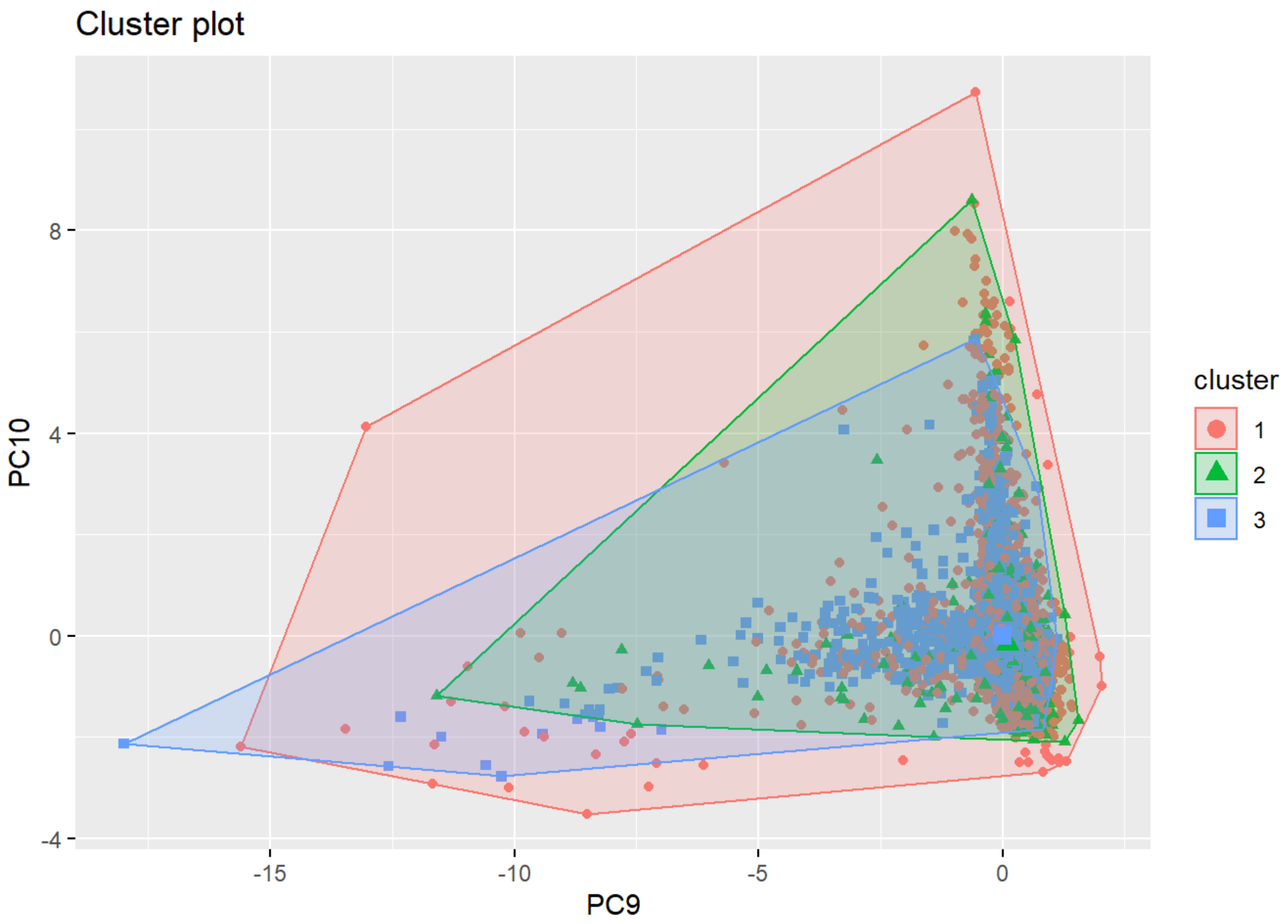
Plotting the clusters overall do not seem very good at 10 clusters. Therefore, I have attempted to reduce dimensionality through principal components analysis.



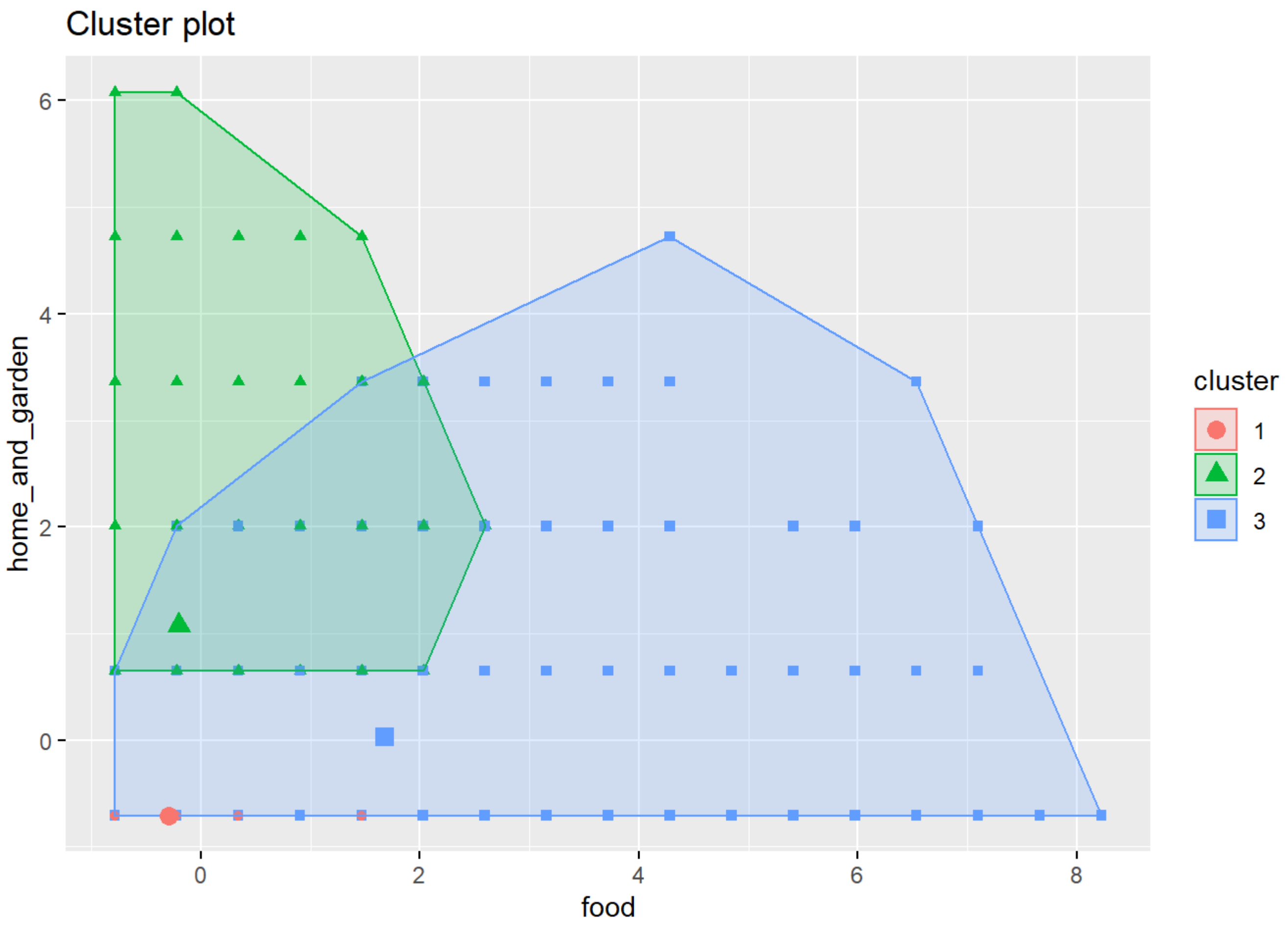
Because the previous cluster did not look too convincing, I have applied principal components analysis to reduce the dimensionality and obtain slightly better results. In the graphic above, the total within-clusters sum of squares has decreased. It appears that the models begin leveling off at a cluster amount of 10, as before. However, the total within-cluster sum of squares is much lower than the data used without principal components analysis.

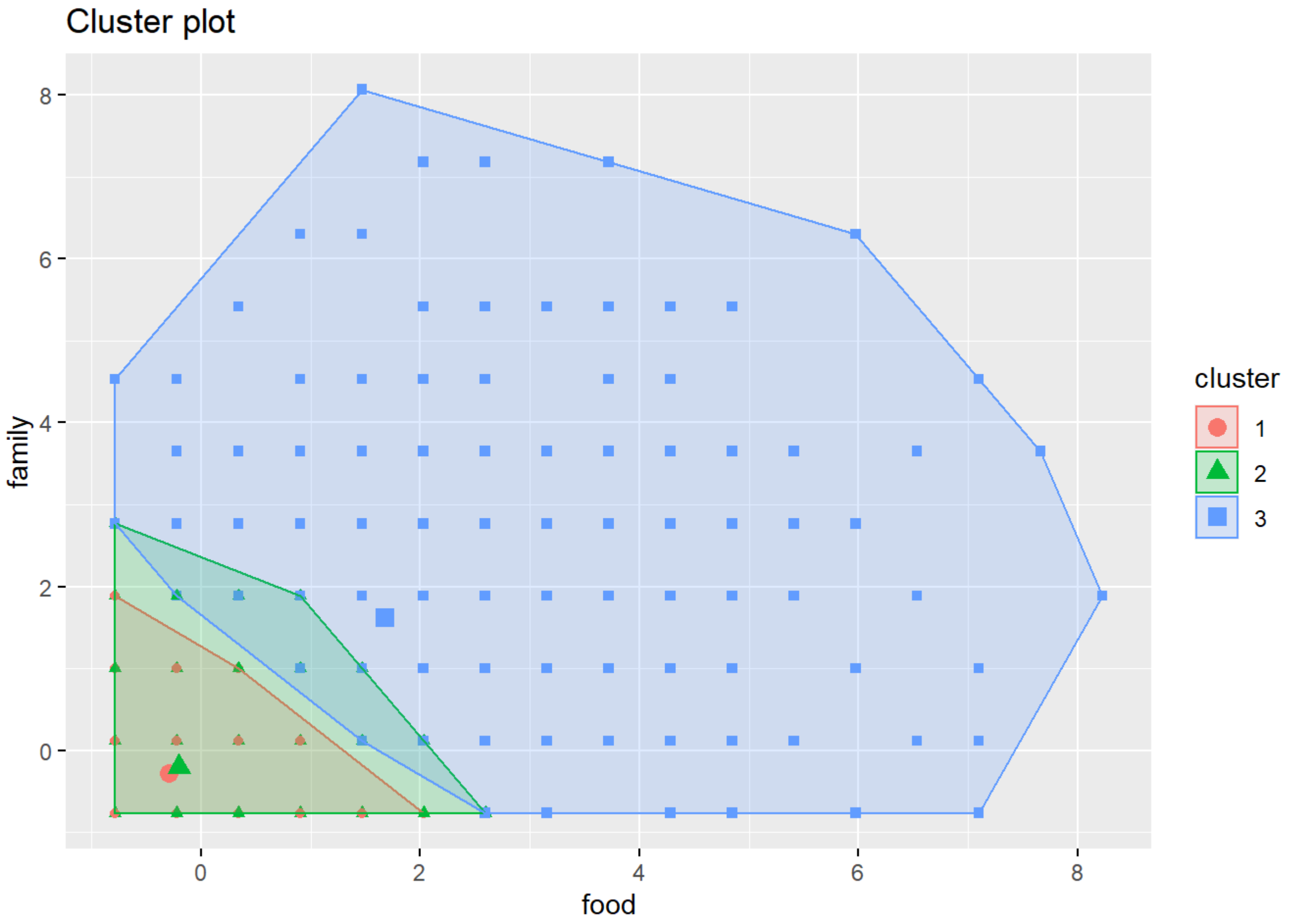


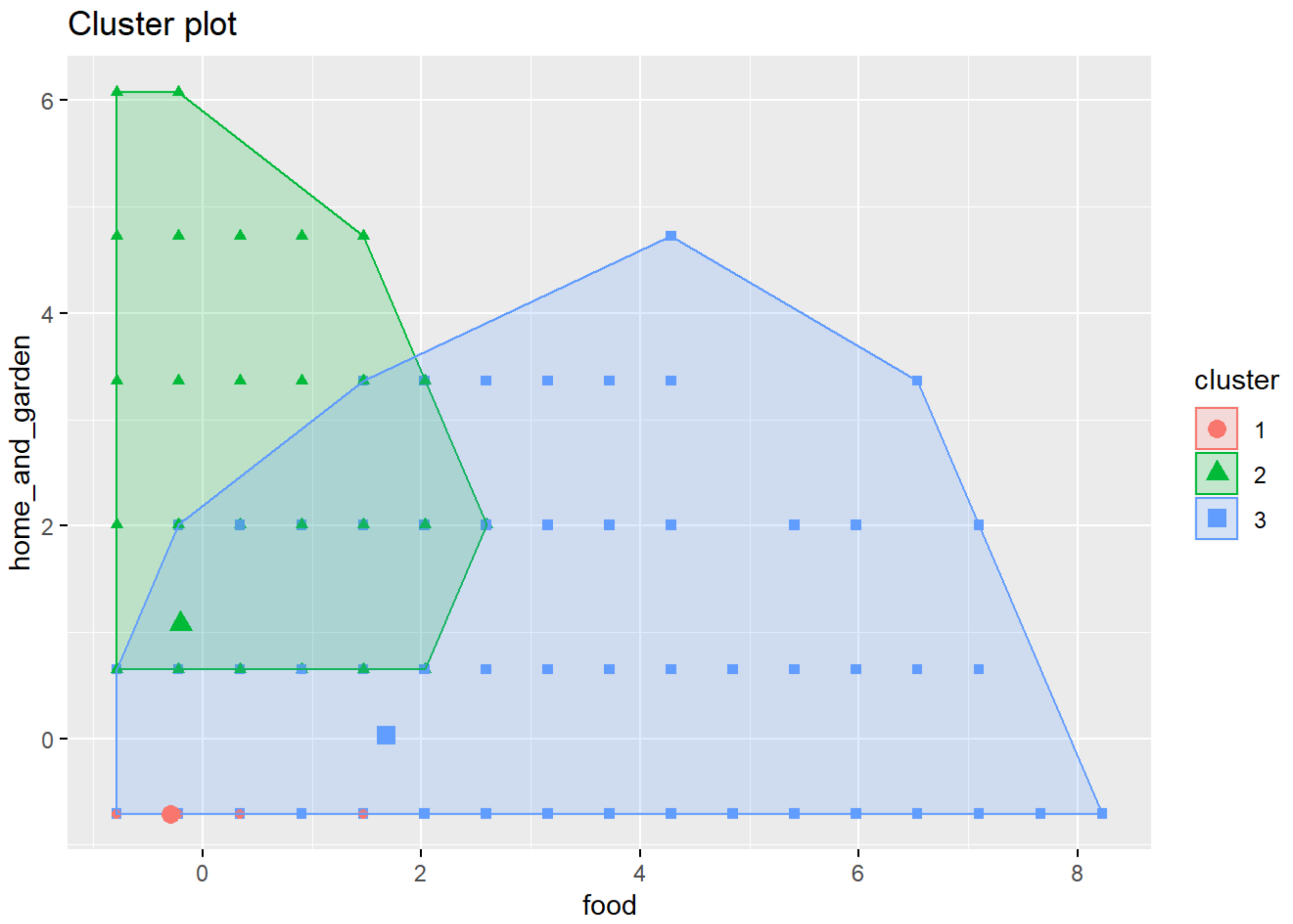
After applying principal components analysis, the clusters seem to look a bit more separated. The previous chart looked very skewed and contained far more overlaps.



However, after attempting to reduce dimensionality, the results do not look convincing. With this understanding, I should attempt to simply use less features in the k-means clustering model.


Cluster plot
home_and_garden
food
6
4
2
0
0
2
4
6
8
cluster
1
2
3


Cluster plot
family
food
8
6
4
2
0
0
2
4
6
8
cluster
1
2
3



I chose to cluster a close market group: those who tweet about food, home and garden, and family. I think attributes such as these could be beneficial for future marketing efforts. Those that like both food and home and garden will likely enjoy organic and nutritional foods. I assume that these individuals are likely tweeting about their gardens and crop yields. The food category was also looked at as these individuals would enjoy food of different kinds. I thought it was also important to add the family category to this cluster analysis. If we can understand the subsets of those who like food, family, and nutritious foods, then we could identify a target market for this product. If I were to understand more about the product itself, then this analysis could be highly tailored to the brand positioning of this product.

#Author attrbution

##Collect raw training data

```
## Warning in tm_map.SimpleCorpus(my_corpus, content_transformer(tolower)):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(my_corpus, content_transformer(removeNumbers)):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(my_corpus,
## content_transformer(removePunctuation)): transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(my_corpus, content_transformer(stripWhitespace)):
## transformation drops documents
## Warning in tm_map.SimpleCorpus(my_corpus, content_transformer(removeWords), :
## transformation drops documents
```

##Collect testing data

```
## Warning in tm_map.SimpleCorpus(my_corpus, content_transformer(tolower)):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(my_corpus, content_transformer(removeNumbers)):
## transformation drops documents
```

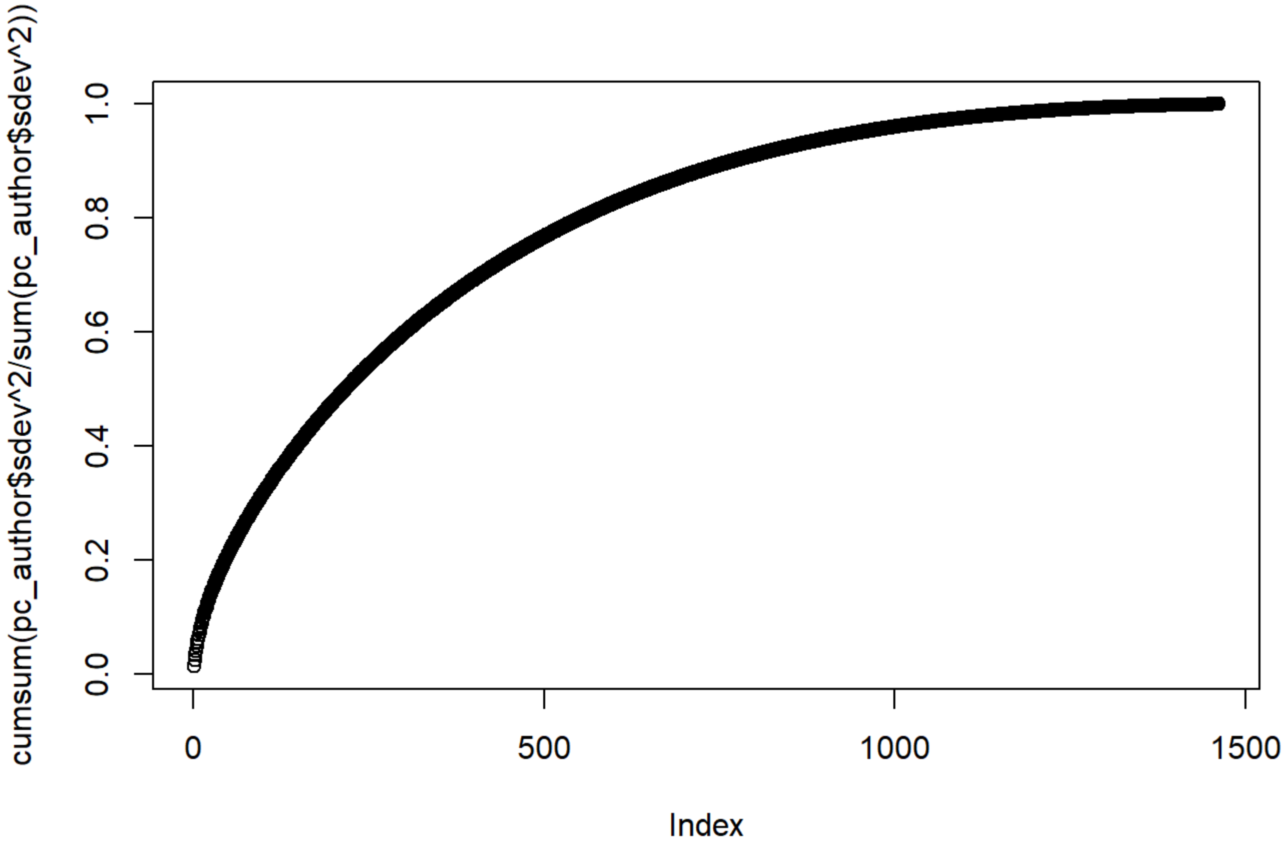
```
## Warning in tm_map.SimpleCorpus(my_corpus,
## content_transformer(removePunctuation)): transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(my_corpus, content_transformer(stripWhitespace)):
## transformation drops documents
```

```
## Warning in tm_map.SimpleCorpus(my_corpus, content_transformer(removeWords), :
## transformation drops documents
```

I assumed that I will simply grab the intersection of shared words between the training and testing sample.

##Principal Components Analysis



```
## [1] 0.516
```

```
## [1] 0.6384
```

This test accuracy of 51% is not very good. I am going to attempt to try a different model.

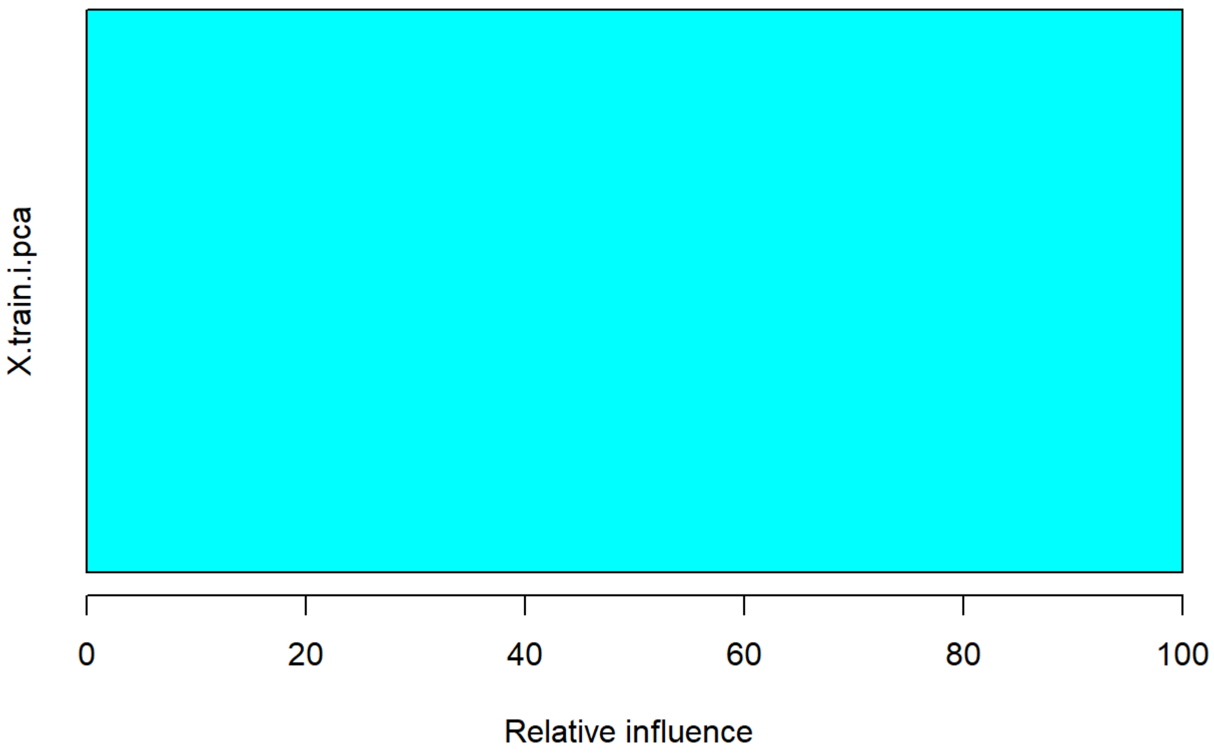
```
##   ObservedAuthor   PredictedAuthor
## 1  AaronPressman     AaronPressman
## 2  AaronPressman     AaronPressman
## 3  AaronPressman      JimGilchrist
## 4  AaronPressman     AaronPressman
## 5  AaronPressman KouroshKarimkhany
## 6  AaronPressman      JimGilchrist
```

```
##   ObservedAuthor   PredictedAuthor
## 1  AaronPressman     AaronPressman
## 2  AaronPressman     AaronPressman
## 3  AaronPressman      JimGilchrist
## 4  AaronPressman     AaronPressman
## 5  AaronPressman KouroshKarimkhany
## 6  AaronPressman      JimGilchrist
```

```
## [1] 0.118
```

The KNN did not run very well. Above, I achieved a 11.8% test accuracy–not good! I played with many different values for the KNN parameters, but did not have great results. This was an attempt at cross validation of model parameters. No luck.

```
## Warning: Setting `distribution = "multinomial"` is ill-advised as it is
## currently broken. It exists only for backwards compatibility. Use at your own
## risk.
```



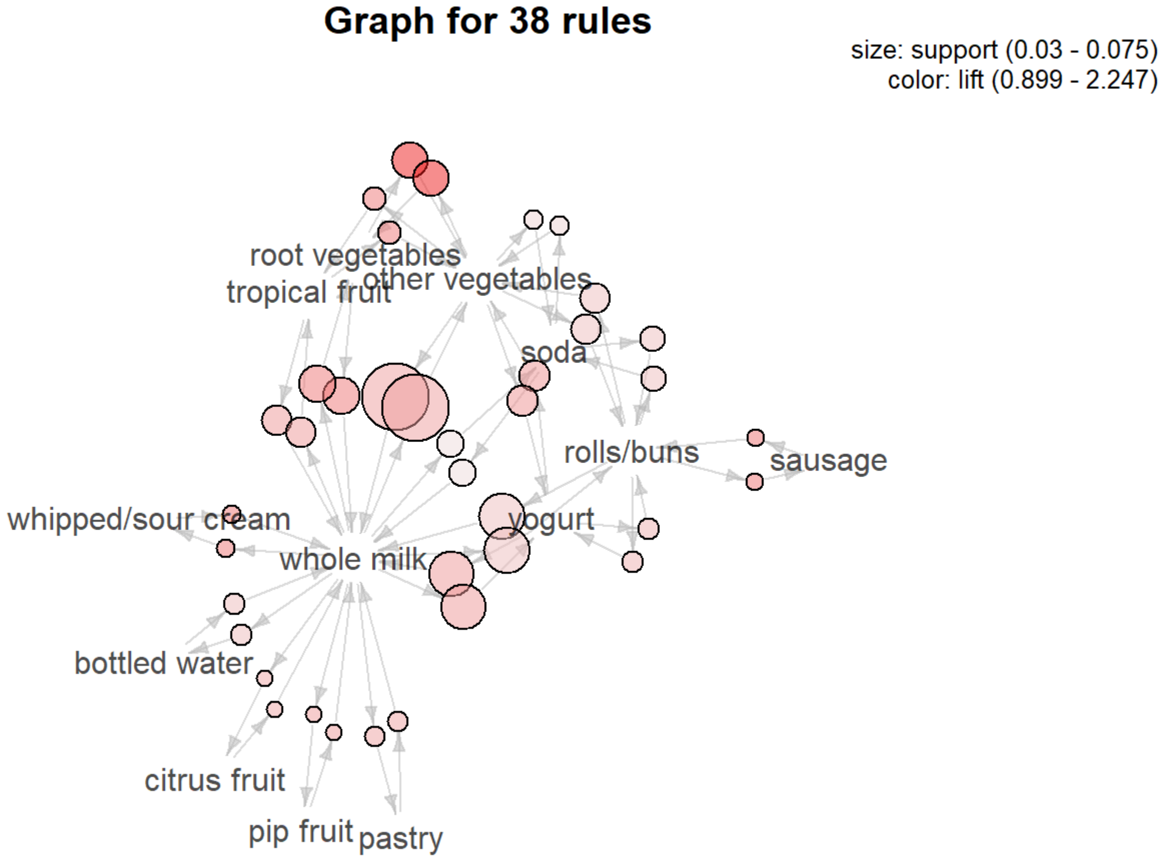
```
##                         var rel.inf
## X.train.i.pca X.train.i.pca     100
```

```
## [1] 0.5352
```

After running a KNN and having little positive result, I ran a boosting forest model. I achieved a slightly higher test accuracy of 52.8%.

#Association rule mining

```
## [1] "citrus fruit,semi-finished bread,margarine,ready soups"
## [2] "tropical fruit,yogurt,coffee"
## [3] "whole milk"
## [4] "pip fruit,yogurt,cream cheese ,meat spreads"
## [5] "other vegetables,whole milk,condensed milk,long life bakery product"
## [6] "whole milk,butter,yogurt,rice,abrasive cleaner"
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##        0.05    0.1    1 none FALSE            TRUE       5    0.03      2
##  maxlen target  ext
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 295
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[169 item(s), 9835 transaction(s)] done [0.01s].
## sorting and recoding items ... [44 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 done [0.00s].
## writing ... [38 rule(s)] done [0.00s].
## creating S4 object  ... done [0.00s].
```



The chart above shows some association rules among common purchases when at the grocery store. Looking at the chart, whole milk is located at the center of many other purchases. To me personally, this makes sense. There are very few times where I would go to the store and not buy milk. Therefore, it is rational to me that whole milk is associated with purchases of many other items. Another interesting relationship is the rolls/buns item and sausage and soda relationships at the left side of the chart. When buying rolls/buns, you are likely buying them to put a sausage inside (at least in my personal life). Nearby is the soda item. To me, soda and these other items seem to go together very well.