Notebook 1

Code ▾

Nguyen Tran

Data: LondonAirBNB (https://www.kaggle.com/datasets/jinxzed/londonairbnb?select=Uphaar_listing.csv)

Hide

```
library(readr)
library(caret)
```

```
Loading required package: ggplot2
Loading required package: lattice
Registered S3 method overwritten by 'data.table':
  method           from
  print.data.table
```

Hide

```
library(tree)
```

Cleaning Data

Hide

```
df <- read.csv("Uphaar_listing.csv")
head(df)
```

	X	price_x	experiences_offered	host_is_superhost	host_listings_count	host_has_profi
	<int>	<int>	<chr>	<chr>	<dbl>	<chr>
1	0	88	family	f	3	t
2	1	65	business	f	4	t
3	2	100	romantic	f	1	t
4	3	300	none	f	18	t
5	4	150	business	f	3	t
6	5	29	none	t	3	t

6 rows | 1-7 of 42 columns

Dropping ‘X’, “city”, “host_location”, “street”, ‘host_since’, “is_business_travel_ready” and “requires_license” and “id”

Hide

```
exclude <- names(df) %in% c("X","city", "host_location", "street", "host_since","is_business_travel_ready", "requires_license", "id")
df <- df[!exclude]
```

There is a need for transforming and cleaning this dataset. There are 41 columns, going through and figuring out which ones are factors

Hide

```
convert.to.numeric <- function(vector) {
  vector <- as.numeric(as.factor(vector))-1
}
```

Looking at the summary of dataframe to consider which ones need to be factorized.

Hide

```
summary(df)
```

```
    price_x        experiences_offered host_is_superhost  host_listings_count host_has_profile_p
ic host_identity_verified is_location_exact  property_type     
 Min.   :    0.0   Length:83709       Length:83709       Min.   :   0.00     Length:83709      
Length:83709           Length:83709       Length:83709      
 1st Qu.:   46.0   Class :character   Class :character   1st Qu.:   1.00     Class :character  
Class :character       Class :character   Class :character  
 Median :   81.0   Mode  :character   Mode  :character   Median :   1.00     Mode  :character  
Mode  :character       Mode  :character   Mode  :character  
 Mean   :  127.5                                         Mean   :  25.97                       
 3rd Qu.:  139.0                                         3rd Qu.:   4.00                       
 Max.   :19970.0                                         Max.   :3142.00                       
                                                         NA's   :11                            
 room_type_y         accommodates      bathrooms         bedrooms            beds          bed_typ
e         guests_included   extra_people     minimum_nights_y
 Length:83709       Min.   : 1.000   Min.   : 0.000   Min.   : 0.000   Min.   : 0.00   Length:83
709       Min.   : 1.000   Min.   :  0.000   Min.   :   1.000  
 Class :character   1st Qu.: 2.000   1st Qu.: 1.000   1st Qu.: 1.000   1st Qu.: 1.00   Class :ch
aracter   1st Qu.: 1.000   1st Qu.:  0.000   1st Qu.:   1.000  
 Mode  :character   Median : 2.000   Median : 1.000   Median : 1.000   Median : 1.00   Mode  :ch
aracter   Median : 1.000   Median :  0.000   Median :   2.000  
                    Mean   : 3.173   Mean   : 1.308   Mean   : 1.409   Mean   : 1.74           
Mean   : 1.607   Mean   :  7.393   Mean   :   4.632  
                    3rd Qu.: 4.000   3rd Qu.: 1.500   3rd Qu.: 2.000   3rd Qu.: 2.00           
3rd Qu.: 2.000   3rd Qu.: 10.000   3rd Qu.:   3.000  
                    Max.   :32.000   Max.   :35.000   Max.   :50.000   Max.   :50.00           
Max.   :46.000   Max.   :247.000   Max.   :1125.000  
                                     NA's   :133      NA's   :164      NA's   :984             
 maximum_nights      availability_30  availability_60 availability_90 availability_365_y number_
of_reviews_y number_of_reviews_ltm instant_bookable   cancellation_policy
 Min.   :        1   Min.   : 0.000   Min.   : 0.00   Min.   : 0.00   Min.   :  0.0      Min.   
:  0.0       Min.   :  0.00        Length:83709       Length:83709      
 1st Qu.:       31   1st Qu.: 0.000   1st Qu.: 0.00   1st Qu.: 0.00   1st Qu.:  0.0      1st Q
u.:  0.0       1st Qu.:  0.00        Class :character   Class :character  
 Median :     1125   Median : 0.000   Median : 7.00   Median :16.00   Median : 69.0      Median 
:  4.0       Median :  1.00        Mode  :character   Mode  :character  
 Mean   :    19156   Mean   : 4.907   Mean   :18.48   Mean   :32.51   Mean   :117.4      Mean   
: 16.9       Mean   :  4.75                                            
 3rd Qu.:     1125   3rd Qu.:11.000   3rd Qu.:40.00   3rd Qu.:70.00   3rd Qu.:249.0      3rd Q
u.: 17.0       3rd Qu.:  5.00                                            
 Max.   :999999999   Max.   :30.000   Max.   :60.00   Max.   :90.00   Max.   :365.0      Max.   
:775.0       Max.   :334.00                                            
                                                                                                 
 require_guest_profile_picture require_guest_phone_verification calculated_host_listings_count_y
calculated_host_listings_count_entire_homes
 Length:83709                  Length:83709                     Min.   :  1.00                  
Min.   :  0.00                             
 Class :character              Class :character                 1st Qu.:  1.00                  
1st Qu.:  0.00                             
 Mode  :character              Mode  :character                 Median :  1.00                  
Median :  1.00                             
                                                                Mean   : 19.59                  
Mean   : 16.76                             
```

```
                                                                3rd Qu.:  4.00
3rd Qu.:  2.00
                                                                Max.   :921.00
Max.   :919.00

 calculated_host_listings_count_private_rooms calculated_host_listings_count_shared_rooms securi
ty_deposit   cleaning_fee
 Min.   :  0.000                             Min.   : 0.00000                            Lengt
h:83709       Length:83709
 1st Qu.:  0.000                             1st Qu.: 0.00000                            Class
:character   Class :character
 Median :  1.000                             Median : 0.00000                            Mode
:character   Mode  :character
 Mean   :  2.166                             Mean   : 0.05133
 3rd Qu.:  1.000                             3rd Qu.: 0.00000
 Max.   :223.000                             Max.   :18.00000
```

Encoding the following variables for data exploration

Hide

```
encodables <- c("experiences_offered","host_is_superhost","host_has_profile_pic","host_identity_
verified",
                "is_location_exact","property_type","room_type_y","bed_type",
                "instant_bookable","cancellation_policy","require_guest_profile_picture",
                "require_guest_phone_verification")

for(col in encodables) {
  df[,col] <- convert.to.numeric(df[,col])
}
```

security_deposit and cleaning_fee are characters currently, representing currency, parsing the numbers out of the column.

Hide

```
df$security_deposit <- parse_number(df$security_deposit)
df$cleaning_fee <- parse_number(df$cleaning_fee)
```

Exploring NA’s in data to see if these can get replaced by the mean of the column.

Hide

```
df$host_listings_count[is.na(df$host_listings_count)] <- mean(df$host_listings_count,na.rm=TRUE)
df$bathrooms[is.na(df$bathrooms)] <- mean(df$bathrooms,na.rm=TRUE)
df$bedrooms[is.na(df$bedrooms)] <- mean(df$bedrooms,na.rm=TRUE)
df$beds[is.na(df$beds)] <- mean(df$beds,na.rm=TRUE)
df$security_deposit[is.na(df$security_deposit)] <- mean(df$security_deposit,na.rm=TRUE)
df$cleaning_fee[is.na(df$cleaning_fee)] <- mean(df$cleaning_fee,na.rm=TRUE)
```

Hide

```
head(df)
```

	price_x	experiences_offered	host_is_superhost	host_listings_count	host_has_prof
	<int>	<dbl>	<dbl>	<dbl>	
1	88	1	1	3	
2	65	0	1	4	
3	100	3	1	1	
4	300	2	1	18	
5	150	0	1	3	
6	29	2	2	3	

6 rows | 1-6 of 34 columns

Train / Test Split

Splitting dataset into the training and test (80%/20%)

Hide

```
set.seed(123) # reproducibility
i <- sample(1:nrow(df), nrow(df)*.80, replace=FALSE)
train <- df[i,]
test <- df[-i,]
```

Hide

```
train
```

	price_x	experiences_offered	host_is_superhost	host_listings_count	host_has_
	<int>	<dbl>	<dbl>	<dbl>	
51663	70	2	1	1	
57870	121	2	2	1	
2986	32	2	1	1	
29925	45	2	1	1	
68293	100	2	1	1	
62555	150	2	1	11	
45404	35	2	2	1	
65161	115	2	1	3	

	price_x	experiences_offered	host_is_superhost	host_listings_count	host_has_
	<int>	<dbl>	<dbl>	<dbl>	
46435	150	2	2	1	
9642	95	2	1	7	

1-10 of 66,967 rows | 1-6 of 34 columns Previous **1** 2 3 4 5 6 ... 100 Next

Exploring Data

Hide

```
suppressWarnings(cor(train))
```

```
                                                   price_x experiences_offered host_is_superhost
host_listings_count host_has_profile_pic host_identity_verified
price_x                                       1.0000000000       -0.0067421142      -0.023725338
0.1663967363        -0.0023581398           -0.042989922
experiences_offered                          -0.0067421142        1.0000000000      -0.013591841
0.0037882267         0.0001125073           -0.022259940
host_is_superhost                            -0.0237253378       -0.0135918415       1.000000000
-0.0610369289         0.0254434617            0.075619634
host_listings_count                           0.1663967363        0.0037882267      -0.061036929
1.0000000000         0.0094878746           -0.077264590
host_has_profile_pic                         -0.0023581398        0.0001125073       0.025443462
0.0094878746         1.0000000000            0.049873378
host_identity_verified                       -0.0429899216       -0.0222599405       0.075619634
-0.0772645895         0.0498733775            1.000000000
is_location_exact                            -0.0018161181       -0.0065166529       0.041912977
-0.0837713851        -0.0032147661            0.014943387
property_type                                 0.0183285892       -0.0163809924       0.057880494
0.0154471708        -0.0042187036           -0.023645723
room_type_y                                  -0.1878753447        0.0216824568       0.053709304
-0.1088322567        -0.0211390195           -0.020696571
accommodates                                  0.2626851733       -0.0302762599      -0.032092339
0.1451887326         0.0173942705            0.007558997
bathrooms                                     0.1997602983       -0.0192410770      -0.019169887
0.1031078657         0.0051406485           -0.007301477
bedrooms                                      0.2263977986       -0.0316807318      -0.021047035
0.1470528796         0.0089424772            0.023539282
beds                                          0.2050605423       -0.0323056777      -0.028332863
0.1360640398         0.0096262958           -0.004299932
bed_type                                      0.0167297082       -0.0108997642      -0.004371366
0.0118860953        -0.0045811445           -0.028053682
guests_included                               0.1873605144       -0.0234546178      -0.006942722
0.2358262620         0.0106906427           -0.012507487
extra_people                                  0.0311587403       -0.0169632430       0.073398571
-0.0510251666         0.0021022809            0.029930218
minimum_nights_y                              0.0366950233       -0.0069164681      -0.018447265
-0.0063416174         0.0054119706            0.022739744
maximum_nights                               -0.0008063852        0.0001244833      -0.001693518
-0.0005758112         0.0002478717            0.005226243
availability_30                               0.0459545355       -0.0155221317       0.077522761
-0.0440034506        -0.0146190805           -0.051512161
availability_60                               0.0640443467       -0.0163488979       0.074222302
-0.0409143392        -0.0123863797           -0.054309446
availability_90                               0.0665370890       -0.0158674982       0.077651153
-0.0412637357        -0.0110135333           -0.056058465
availability_365_y                            0.0900999948       -0.0249889188       0.052345237
-0.0294376932        -0.0091428997           -0.040989984
number_of_reviews_y                          -0.0548052080       -0.0115775966       0.281531438
-0.0539368607         0.0179156794            0.117802987
number_of_reviews_ltm                        -0.0450640008        0.0041920273       0.311983592
-0.0513843938         0.0193122782            0.002977735
instant_bookable                              0.0592326998        0.0143609603      -0.049372508
0.1345013216        -0.0108195040           -0.158831547
```

```
cancellation_policy                          0.0579761799      -0.0261248513       0.087062193
0.1351874698         0.0214099724           0.069122885
require_guest_profile_picture               -0.0104521136      -0.0392836555       0.045209655
-0.0124641879         0.0061048773           0.062099885
require_guest_phone_verification             0.0014957474      -0.0326306459       0.032509003
0.0136455316         0.0078166858           0.080090033
calculated_host_listings_count_y             0.1502532651       0.0036273812      -0.063426146
0.8947683689         0.0107833787          -0.074109121
                                            is_location_exact property_type  room_type_y  accom
modates    bathrooms     bedrooms         beds      bed_type guests_included
price_x                                         -0.0018161181   0.018328589 -0.187875345  0.262
6851733  0.199760298  0.226397799  0.205060542  1.672971e-02     0.187360514
experiences_offered                             -0.0065166529  -0.016380992  0.021682457 -0.030
2762599 -0.019241077 -0.031680732 -0.032305678 -1.089976e-02    -0.023454618
host_is_superhost                                0.0419129775   0.057880494  0.053709304 -0.032
0923392 -0.019169887 -0.021047035 -0.028332863 -4.371366e-03    -0.006942722
host_listings_count                             -0.0837713851   0.015447171 -0.108832257  0.145
1887326  0.103107866  0.147052880  0.136064040  1.188610e-02     0.235826262
host_has_profile_pic                            -0.0032147661  -0.004218704 -0.021139019  0.017
3942705  0.005140649  0.008942477  0.009626296 -4.581145e-03     0.010690643
host_identity_verified                           0.0149433874  -0.023645723 -0.020696571  0.007
5589971 -0.007301477  0.023539282 -0.004299932 -2.805368e-02    -0.012507487
is_location_exact                                1.0000000000   0.073300680 -0.017510945  0.013
0582370  0.029325593  0.023477754  0.013535443 -9.019503e-04    -0.018094455
property_type                                    0.0733006803   1.000000000  0.217591182  0.065
5431292  0.205495227  0.163531988  0.121286614  2.697069e-03     0.026219218
room_type_y                                     -0.0175109451   0.217591182  1.000000000 -0.537
6147146 -0.154993532 -0.352460682 -0.372537508 -3.539304e-02    -0.303098020
accommodates                                     0.0130582370   0.065543129 -0.537614715  1.000
0000000  0.486517088  0.751759714  0.794293606  3.061522e-02     0.529051084
bathrooms                                        0.0293255929   0.205495227 -0.154993532  0.486
5170883  1.000000000  0.593577019  0.495174168  2.272909e-02     0.274158751
bedrooms                                         0.0234777542   0.163531988 -0.352460682  0.751
7597142  0.593577019  1.000000000  0.722143064  3.219163e-02     0.416555791
beds                                             0.0135354433   0.121286614 -0.372537508  0.794
2936057  0.495174168  0.722143064  1.000000000  3.283102e-02     0.442031176
bed_type                                        -0.0009019503   0.002697069 -0.035393038  0.030
6152239  0.022729087  0.032191634  0.032831022  1.000000e+00     0.021862918
guests_included                                 -0.0180944550   0.026219218 -0.303098020  0.529
0510842  0.274158751  0.416555791  0.442031176  2.186292e-02     1.000000000
extra_people                                    -0.0052580198   0.019859595 -0.042687446  0.116
1533701  0.051567523  0.072810208  0.088622615  2.338228e-03     0.268640298
minimum_nights_y                                -0.0058476800  -0.013168057 -0.029519958 -0.000
3416719  0.006846080  0.009880198  0.002098955  1.632088e-03    -0.003270478
maximum_nights                                   0.0025683656   0.005012831  0.004435782 -0.004
2590723 -0.001901828 -0.001730709 -0.005275439  3.000275e-04    -0.001751121
availability_30                                 -0.0045418263   0.061236012  0.094262092 -0.008
2529385  0.009033050 -0.040767020 -0.008478052  4.823172e-03     0.009825278
availability_60                                 -0.0039009298   0.060674046  0.073750964  0.014
2044100  0.023713339 -0.027597698  0.011346975  6.859145e-03     0.025689867
availability_90                                 -0.0025626508   0.061408211  0.072191786  0.015
8394651  0.023884377 -0.030254131  0.012421278  7.013615e-03     0.026729806
```

```
availability_365_y                                0.0034213109   0.068719825  0.018553924  0.070
3291183  0.041432897 -0.005018585  0.054693812  1.617300e-03     0.066291758
number_of_reviews_y                               0.0299768328   0.018618854  0.090696747 -0.032
6508379 -0.037129423 -0.061823189 -0.033886359 -2.561217e-02     0.013690664
number_of_reviews_ltm                             0.0282028406   0.016109059  0.079983806 -0.015
5876843 -0.027195271 -0.062819675 -0.014822458  1.495725e-03     0.022287644
instant_bookable                                 -0.0151079032   0.010826657  0.025161029  0.024
8734292 -0.004002623 -0.047745818  0.016869953  3.694010e-02     0.038473629
cancellation_policy                              -0.0097427025  -0.055946095 -0.239535227  0.205
1923609  0.071748526  0.123627874  0.144365997  1.207120e-02     0.200026045
require_guest_profile_picture                     0.0235835745   0.016214402  0.008370418 -0.011
0124247 -0.007716830 -0.007968078 -0.009644394 -8.713657e-03    -0.003792267
require_guest_phone_verification                  0.0171529920   0.004338892 -0.024856160  0.023
1053491  0.001157264  0.012310992  0.014271905 -3.282141e-03     0.054040220
calculated_host_listings_count_y                 -0.0863272705   0.023986955 -0.122102038  0.167
5737308  0.118659835  0.167682069  0.152632531  1.286130e-02     0.265446265
                                             extra_people minimum_nights_y maximum_nights availa
bility_30 availability_60 availability_90 availability_365_y
price_x                                       0.031158740     0.0366950233  -8.063852e-04     0.
045954535     0.064044347     0.066537089        0.090099995
experiences_offered                          -0.016963243    -0.0069164681   1.244833e-04    -0.
015522132    -0.016348898    -0.015867498       -0.024988919
host_is_superhost                             0.073398571    -0.0184472653  -1.693518e-03     0.
077522761     0.074222302     0.077651153        0.052345237
host_listings_count                          -0.051025167    -0.0063416174  -5.758112e-04    -0.
044003451    -0.040914339    -0.041263736       -0.029437693
host_has_profile_pic                          0.002102281     0.0054119706   2.478717e-04    -0.
014619081    -0.012386380    -0.011013533       -0.009142900
host_identity_verified                        0.029930218     0.0227397441   5.226243e-03    -0.
051512161    -0.054309446    -0.056058465       -0.040989984
is_location_exact                            -0.005258020    -0.0058476800   2.568366e-03    -0.
004541826    -0.003900930    -0.002562651        0.003421311
property_type                                 0.019859595    -0.0131680567   5.012831e-03     0.
061236012     0.060674046     0.061408211        0.068719825
room_type_y                                  -0.042687446    -0.0295199576   4.435782e-03     0.
094262092     0.073750964     0.072191786        0.018553924
accommodates                                  0.116153370    -0.0003416719  -4.259072e-03    -0.
008252939     0.014204410     0.015839465        0.070329118
bathrooms                                     0.051567523     0.0068460801  -1.901828e-03     0.
009033050     0.023713339     0.023884377        0.041432897
bedrooms                                      0.072810208     0.0098801978  -1.730709e-03    -0.
040767020    -0.027597698    -0.030254131       -0.005018585
beds                                          0.088622615     0.0020989551  -5.275439e-03    -0.
008478052     0.011346975     0.012421278        0.054693812
bed_type                                      0.002338228     0.0016320881   3.000275e-04     0.
004823172     0.006859145     0.007013615        0.001617300
guests_included                               0.268640298    -0.0032704778  -1.751121e-03     0.
009825278     0.025689867     0.026729806        0.066291758
extra_people                                  1.000000000    -0.0062577761  -2.066431e-03     0.
079421716     0.086174072     0.087381977        0.084788485
minimum_nights_y                             -0.006257776     1.0000000000  -7.061842e-04     0.
027565507     0.033597746     0.034279107        0.041923314
```

```
maximum_nights                             -0.002066431   -0.0007061842   1.000000e+00     0.
005698776     0.004900163     0.004738915       0.006478851
availability_30                             0.079421716    0.0275655065   5.698776e-03     1.
000000000     0.954985675     0.925042654       0.662534897
availability_60                             0.086174072    0.0335977457   4.900163e-03     0.
954985675     1.000000000     0.986993963       0.716551724
availability_90                             0.087381977    0.0342791073   4.738915e-03     0.
925042654     0.986993963     1.000000000       0.743495929
availability_365_y                          0.084788485    0.0419233143   6.478851e-03     0.
662534897     0.716551724     0.743495929       1.000000000
number_of_reviews_y                         0.067322565   -0.0301911036   1.396820e-02     0.
076568736     0.074005708     0.079618657       0.107476559
number_of_reviews_ltm                       0.038047263   -0.0462768680   7.263806e-03     0.
110842422     0.112603996     0.120466347       0.111328731
instant_bookable                           -0.034836777   -0.0168357537   4.658295e-03     0.
039961891     0.045553073     0.047545922       0.068907068
cancellation_policy                         0.117390401    0.0127942271   3.874040e-03     0.
049093767     0.063751498     0.067920615       0.106420077
require_guest_profile_picture               0.026003088    0.0070092653  -3.999833e-04     0.
035470324     0.037767741     0.038265740       0.051954247
require_guest_phone_verification            0.041433971    0.0030292926  -5.045750e-04     0.
033773665     0.039225006     0.041211139       0.065706532
calculated_host_listings_count_y           -0.054782179   -0.0049880692  -6.375886e-04    -0.
046796651    -0.043815834    -0.045836557      -0.024873745
                                           number_of_reviews_y number_of_reviews_ltm instant_b
ookable cancellation_policy require_guest_profile_picture
price_x                                          -0.0548052080         -0.045064001      0.05
9232700         0.057976180                 -0.0104521136
experiences_offered                              -0.0115775966          0.004192027      0.01
4360960        -0.026124851                 -0.0392836555
host_is_superhost                                 0.2815314380          0.311983592     -0.04
9372508         0.087062193                  0.0452096549
host_listings_count                              -0.0539368607         -0.051384394      0.13
4501322         0.135187470                 -0.0124641879
host_has_profile_pic                              0.0179156794          0.019312278     -0.01
0819504         0.021409972                  0.0061048773
host_identity_verified                            0.1178029867          0.002977735     -0.15
8831547         0.069122885                  0.0620998852
is_location_exact                                 0.0299768328          0.028202841     -0.01
5107903        -0.009742702                  0.0235835745
property_type                                     0.0186188542          0.016109059      0.01
0826657        -0.055946095                  0.0162144022
room_type_y                                       0.0906967471          0.079983806      0.02
5161029        -0.239535227                  0.0083704177
accommodates                                     -0.0326508379         -0.015587684      0.02
4873429         0.205192361                 -0.0110124247
bathrooms                                        -0.0371294230         -0.027195271     -0.00
4002623         0.071748526                 -0.0077168301
bedrooms                                         -0.0618231895         -0.062819675     -0.04
7745818         0.123627874                 -0.0079680784
beds                                             -0.0338863591         -0.014822458      0.01
6869953         0.144365997                 -0.0096443943
```

```
bed_type                                           -0.0256121679           0.001495725      0.03
6940096         0.012071199                 -0.0087136571
guests_included                                     0.0136906635           0.022287644      0.03
8473629         0.200026045                 -0.0037922665
extra_people                                        0.0673225654           0.038047263     -0.03
4836777         0.117390401                  0.0260030883
minimum_nights_y                                   -0.0301911036          -0.046276868     -0.01
6835754         0.012794227                  0.0070092653
maximum_nights                                      0.0139681972           0.007263806      0.00
4658295         0.003874040                 -0.0003999833
availability_30                                     0.0765687355           0.110842422      0.03
9961891         0.049093767                  0.0354703240
availability_60                                     0.0740057085           0.112603996      0.04
5553073         0.063751498                  0.0377677412
availability_90                                     0.0796186571           0.120466347      0.04
7545922         0.067920615                  0.0382657402
availability_365_y                                  0.1074765591           0.111328731      0.06
8907068         0.106420077                  0.0519542474
number_of_reviews_y                                 1.0000000000           0.692967926      0.00
6215966         0.152924287                  0.1114144316
number_of_reviews_ltm                               0.6929679255           1.000000000      0.09
4789048         0.138186612                  0.0222186523
instant_bookable                                    0.0062159658           0.094789048      1.00
0000000        -0.030585996                 -0.0375870505
cancellation_policy                                 0.1529242872           0.138186612     -0.03
0585996         1.000000000                  0.0441067037
require_guest_profile_picture                       0.1114144316           0.022218652     -0.03
7587050         0.044106704                  1.0000000000
require_guest_phone_verification                    0.1078535126           0.022100630     -0.01
0574076         0.053326267                  0.6651862210
calculated_host_listings_count_y                   -0.0557015529          -0.051839822      0.13
2807680         0.138198851                 -0.0133733088
                                             require_guest_phone_verification calculated_host_li
stings_count_y calculated_host_listings_count_entire_homes
price_x                                                           0.001495747
0.1502532651                                0.1547197068
experiences_offered                                              -0.032630646
0.0036273812                                0.0031979067
host_is_superhost                                                 0.032509003
-0.0634261458                               -0.0581317082
host_listings_count                                               0.013645532
0.8947683689                                0.8879796742
host_has_profile_pic                                              0.007816686
0.0107833787                                0.0102653869
host_identity_verified                                            0.080090033
-0.0741091212                               -0.0656839372
is_location_exact                                                 0.017152992
-0.0863272705                               -0.0771443039
property_type                                                     0.004338892
0.0239869549                                0.0187570718
room_type_y                                                      -0.024856160
-0.1221020383                               -0.1409824965
```

```
accommodates                                                      0.023105349
0.1675737308                                0.1727063060
bathrooms                                                         0.001157264
0.1186598350                                0.1232185973
bedrooms                                                          0.012310992
0.1676820689                                0.1777484996
beds                                                              0.014271905
0.1526325309                                0.1554511478
bed_type                                                         -0.003282141
0.0128612953                                0.0117577245
guests_included                                                   0.054040220
0.2654462649                                0.2671990731
extra_people                                                      0.041433971
-0.0547821787                               -0.0558449391
minimum_nights_y                                                  0.003029293
-0.0049880692                               -0.0044028222
maximum_nights                                                   -0.000504575
-0.0006375886                               -0.0006556852
availability_30                                                   0.033773665
-0.0467966512                               -0.0620408910
availability_60                                                   0.039225006
-0.0438158341                               -0.0611169691
availability_90                                                   0.041211139
-0.0458365570                               -0.0637956369
availability_365_y                                                0.065706532
-0.0248737450                               -0.0485108688
number_of_reviews_y                                               0.107853513
-0.0557015529                               -0.0551362250
number_of_reviews_ltm                                             0.022100630
-0.0518398220                               -0.0536698062
instant_bookable                                                 -0.010574076
0.1328076803                                0.1295002736
cancellation_policy                                               0.053326267
0.1381988509                                0.1310447785
require_guest_profile_picture                                     0.665186221
-0.0133733088                               -0.0131087550
require_guest_phone_verification                                  1.000000000
0.0235807616                                0.0251866599
calculated_host_listings_count_y                                  0.023580762
1.0000000000                                0.9880298464
                                             calculated_host_listings_count_private_rooms calcul
ated_host_listings_count_shared_rooms security_deposit cleaning_fee
price_x                                                                     -0.0228103656
2.909236e-02     1.345538e-01  0.247801934
experiences_offered                                                          0.0020984474
3.314690e-03    -2.503041e-02 -0.013506432
host_is_superhost                                                           -0.0278330839
-2.782984e-02     2.595129e-03 -0.053853458
host_listings_count                                                          0.1063719011
-1.388643e-03     1.017570e-01  0.400934999
host_has_profile_pic                                                         0.0026636636
3.572832e-03    -5.068144e-03  0.009669859
```

```
host_identity_verified                                                      -0.0473734478
-2.592674e-02     4.078645e-02 -0.052788785
is_location_exact                                                           -0.0516114357
3.713586e-03    -5.860764e-03 -0.012274697
property_type                                                                0.0019882620
2.961655e-02     2.545005e-02  0.018997741
room_type_y                                                                  0.1303234612
9.589852e-02    -1.500926e-01 -0.359718845
accommodates                                                                -0.0164297090
2.219670e-02     1.760922e-01  0.450702969
bathrooms                                                                   -0.0113070672
7.196242e-02     1.686900e-01  0.336216655
bedrooms                                                                    -0.0202891303
-3.110982e-02     1.932609e-01  0.444869230
beds                                                                        -0.0023715701
9.398493e-02     1.559637e-01  0.387562811
bed_type                                                                     0.0069817699
8.560413e-05     2.312033e-03  0.022767270
guests_included                                                              0.0312538843
-2.160267e-02     1.086775e-01  0.384347320
extra_people                                                                 0.0120475167
-2.301364e-03     6.366417e-02  0.049926735
minimum_nights_y                                                            -0.0058882417
-8.208684e-03     5.598363e-02  0.037140531
maximum_nights                                                               0.0002487684
-2.938229e-04     2.586969e-06 -0.003971972
availability_30                                                              0.0881804032
1.582381e-02     1.417277e-02 -0.016071589
availability_60                                                              0.0912516903
3.194194e-02     2.238228e-02 -0.001001482
availability_90                                                              0.0920610699
3.785224e-02     2.198503e-02 -0.002434899
availability_365_y                                                           0.1244479326
5.511872e-02     3.242523e-02  0.014480750
number_of_reviews_y                                                          0.0022693517
1.882165e-04    -4.831156e-02 -0.096702746
number_of_reviews_ltm                                                        0.0162194111
1.925436e-02    -7.482439e-02 -0.095994666
instant_bookable                                                            -0.0070298173
3.681745e-02    -5.234460e-02  0.020718684
cancellation_policy                                                          0.0677996162
-2.862394e-02     7.828071e-02  0.126077038
require_guest_profile_picture                                                0.0011551373
-6.267512e-03     2.492514e-02 -0.016174696
require_guest_phone_verification                                            -0.0067671540
-8.284440e-03     3.939727e-02  0.026755340
calculated_host_listings_count_y                                             0.1464385574
3.801238e-03     1.220339e-01  0.446507846
 [ reached getOption("max.print") -- omitted 5 rows ]
```

It appears the most correlated variables are: “host_listings_count”, “room_type_y”, “accomodates”, “bathrooms”, “bedrooms”, “beds”, “guests_included”, “calculated_host_listings_count_y”, “calculated_host_listings_count_entire_homes”, “security_deposit”, “cleaning_fee” relative to the target price_x.

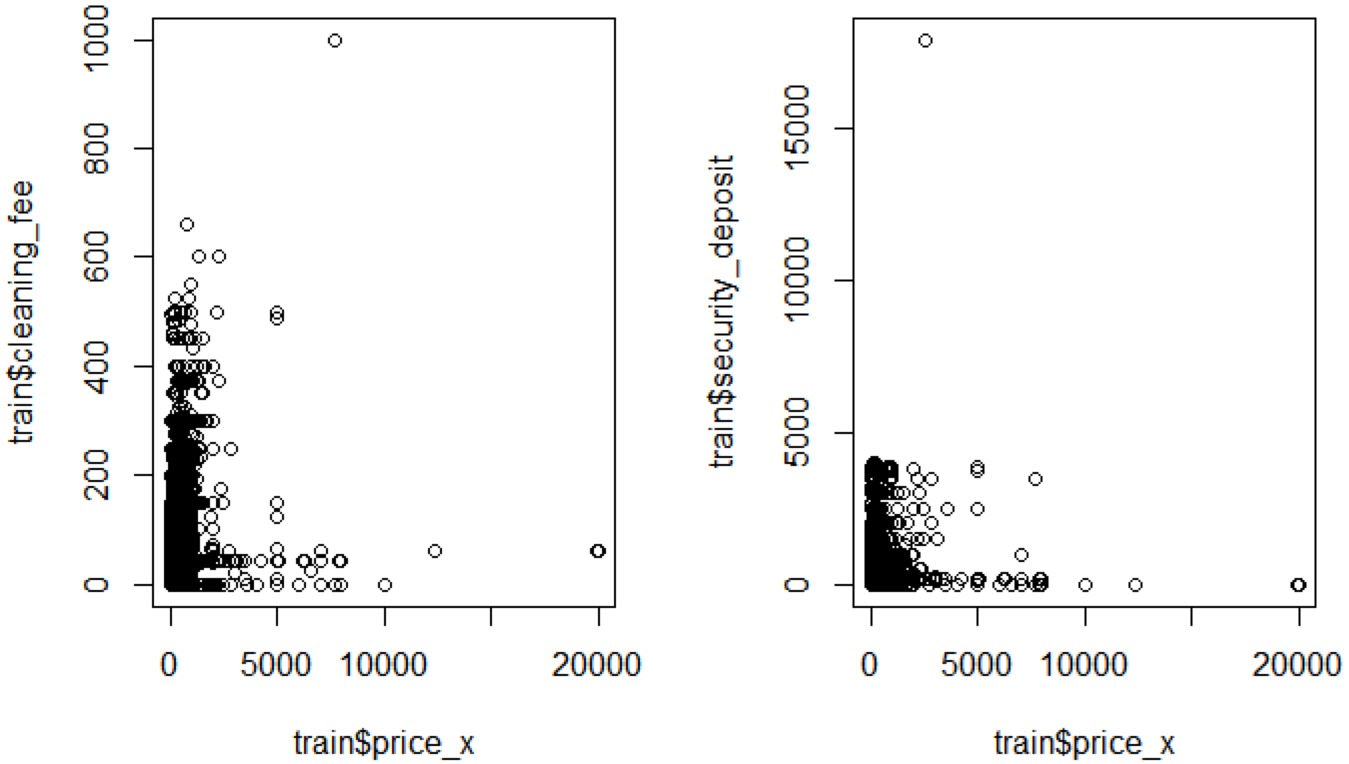
It appears cleaning fees and security deposits are associated with lower range prices for the air bnb.

Hide

```
par(mfrow=c(1,2))
plot(train$price_x, train$cleaning_fee)
plot(train$price_x, train$security_deposit)
```

Hide

```
par(mfrow=c(1,2))
```

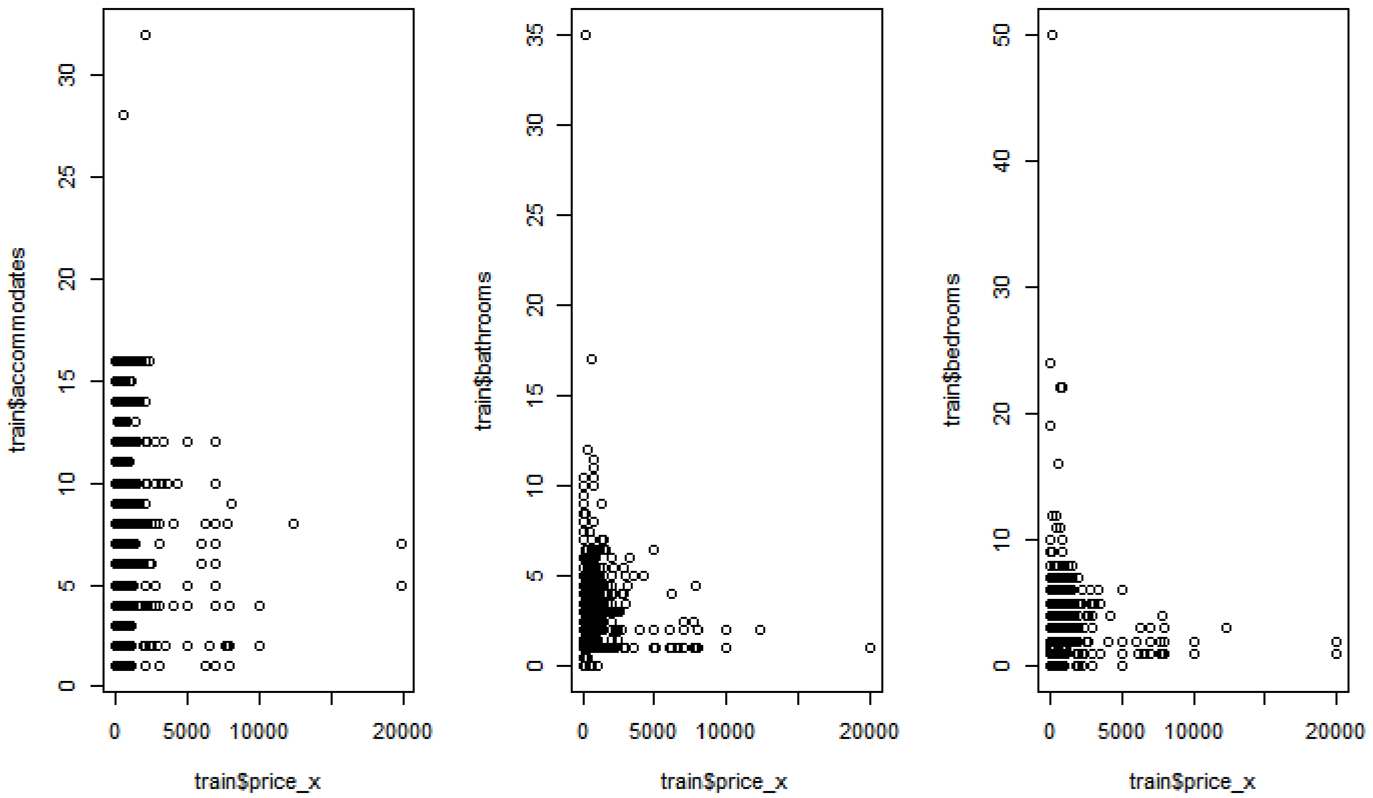


Hide

```
par(mfrow=c(1,3))
plot(train$price_x, train$accommodates)
plot(train$price_x, train$bathrooms)
```

Hide

```
plot(train$price_x, train$bedrooms)
par(mfrow=c(1,3))
```



It appears most of the data tends to cluster around a relatively low price point per night, accommodating 15 or less, with around 5 or less bathrooms and less than 10 bedrooms. Interestingly, there are a few outliers, unbelievable ones really. Such as the mysteriously low-priced per night stay at a place with 50 bedrooms and around 35 bathrooms, accomodating over 30 people. I'd like to have that for myself to be honest haha. What a deal!

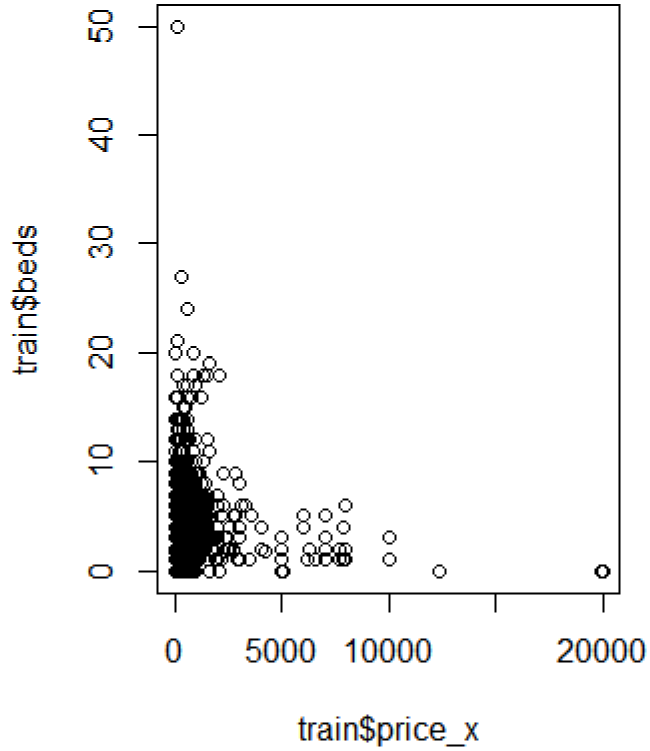
Nonetheless, let's consider more datapoints to get a better view of what the data is telling us.

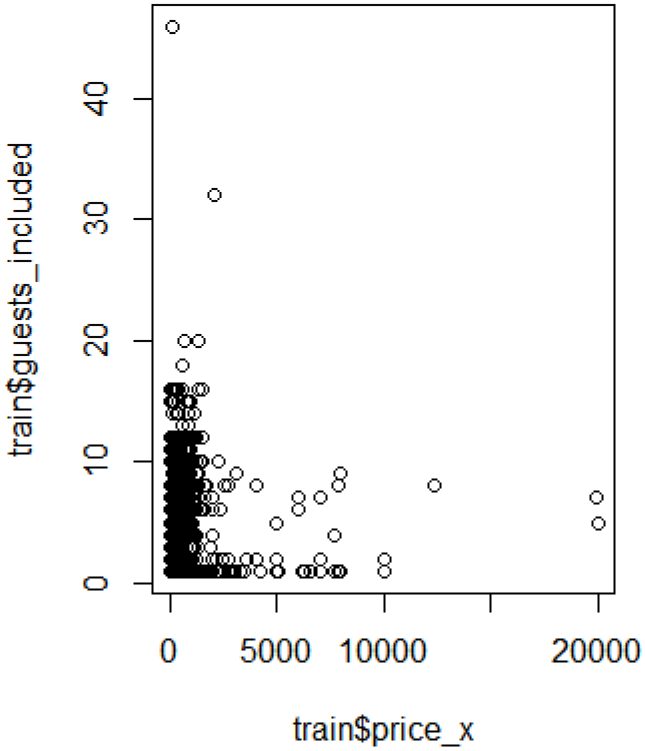
Hide

```
par(mfrow=c(1,2))
plot(train$price_x, train$beds)
plot(train$price_x, train$guests_included)
```

Hide

```
par(mfrow=c(1,2))
```





So it appears other points tend to cluster in those same areas suggesting the same as implied before.

Linear Regression

Hide

```
mod <- lm(price_x ~ host_listings_count + accommodates + bathrooms + bedrooms + beds + guests_included, data=train)
summary(mod)
```

```

Call:
lm(formula = price_x ~ host_listings_count + accommodates + bathrooms + 
    bedrooms + beds + guests_included, data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-1373.9   -49.3   -23.9     8.2 19759.4 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -19.563239   2.440795  -8.015 1.12e-15 ***
host_listings_count   0.228032   0.007092  32.152  < 2e-16 ***
accommodates         27.788278   0.943777  29.444  < 2e-16 ***
bathrooms            38.502405   1.993199  19.317  < 2e-16 ***
bedrooms              6.562241   1.813960   3.618 0.000298 ***
beds                -11.136999   1.350894  -8.244  < 2e-16 ***
guests_included       7.778748   0.876780   8.872  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 258.5 on 66960 degrees of freedom
Multiple R-squared:  0.09379,	Adjusted R-squared:  0.09371 
F-statistic:  1155 on 6 and 66960 DF,  p-value: < 2.2e-16
```

It appears all of the independent variables are statistically significant, yet the R^2 is pretty low, nearing 9%. Let's see how well the model is predicting with respect to correlation and mean squared error.

Hide

```
pred <- predict(mod, newdata=test)
cor_lm <- cor(pred, test$price_x)
mse_lm <- mean((pred-test$price_x)^2)
print(paste("cor=", cor_lm))
```

```
[1] "cor= 0.272380549096964"
```

Hide

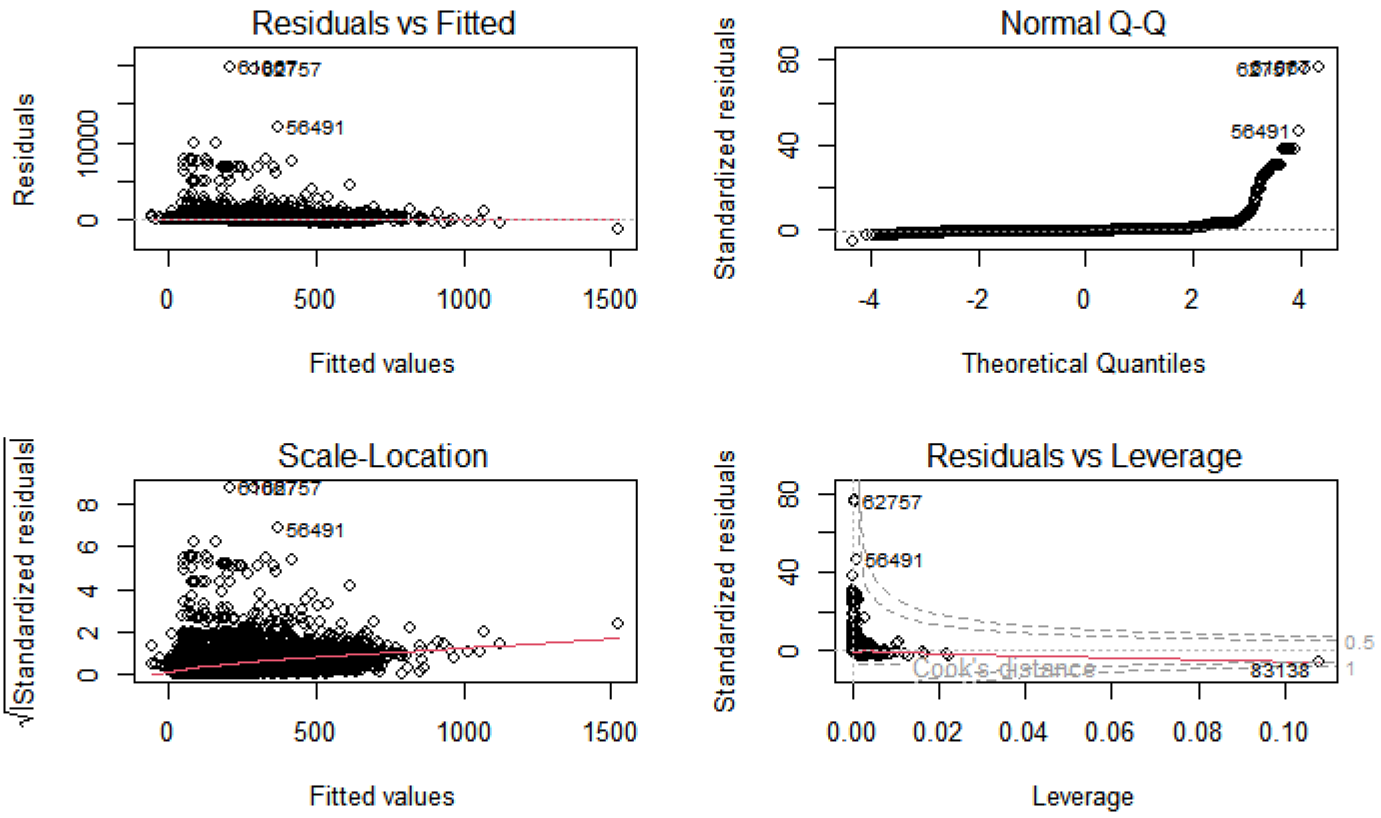
```
print(paste("mse=", mse_lm))
```

```
[1] "mse= 95870.6811589135"
```

Darn. The correlation is low and the error is high. Let's inspect the model performance using residuals plot.

Hide

```
par(mfrow=c(2,2))
plot(mod)
par(mfrow=c(2,2))
```



It appears there's a good number of outliers sitting aside the line in the residuals vs fitted plot, the variance does not appear to be constant, as an assumption of linear model. In the scale-location plot this is seen as well, most of the points appear to cluster around 1 and extend to around 2, however, the number of outliers seem to point at a non-linear relationship being present in the data. In addition the normal q-q points that while most of the data fits to what appears to be linear, the end tail towards the upper theoretical quantile seems to skew away. It appears some of those outliers have high residual, indicating their poor fit by the model.

So it doesn't look like the relationship between these points and the data can be prescribed to a linear regression, let's explore the other kinds of regression.

KNN Regression

We will explore KNN regression using the same variables used in the linear regression: price_x, host_listings_count, accommodates, bathrooms, bedrooms, beds, and guests_included

Hide

```
set.seed(124) # reproducibility
data <- subset(df, select = c(price_x, host_listings_count, accommodates, bathrooms, bedrooms, b
eds, guests_included))
i <- sample(1:nrow(data), nrow(data)*.80, replace=FALSE)
train <- data[i,]
test <- data[-i,]
```

Clustering algorithms work best when the data is scaled. Let's scale the data.

Hide

```
means <- sapply(train, mean)
stdevs <- sapply(train, sd)
train_scaled <- scale(train, center=means, scale=stdevs)
test_scaled <- scale(test, center=means, scale=stdevs)
```

Let’s test various values of k to find the best value for k.

Hide

```
cor_k <- rep(0,20)
mse_k <- rep(0,20)
i <- 1
for(k in seq(1,39,2)) {
  fit_k <- knnreg(train_scaled, train$price_x, k=k)
  pred_k <- predict(fit_k, test_scaled)
  cor_k[i] <- cor(pred_k, test$price_x)
  mse_k[i] <- mean((pred_k - test$price_x)^2)
  print(paste("k=",k,cor_k[i],mse_k[i]))
  i <- i+1
}
```

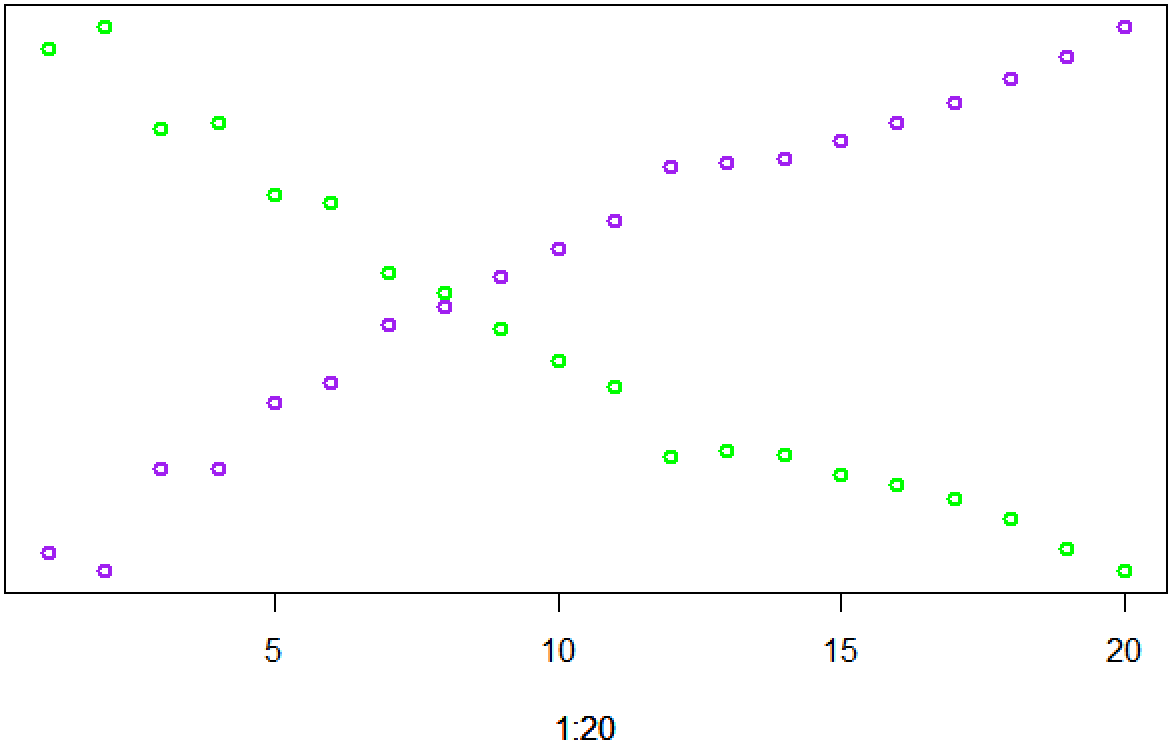
```
[1] "k= 1 0.99289845218099 1335.26132077423"
[1] "k= 3 0.994586889110888 1021.90929102559"
[1] "k= 5 0.987103422856028 2780.47789242571"
[1] "k= 7 0.987463378553037 2771.71592311804"
[1] "k= 9 0.98228937076861 3910.34731044692"
[1] "k= 11 0.981623168707515 4244.56988157346"
[1] "k= 13 0.976617724836065 5249.4019821889"
[1] "k= 15 0.975091605365138 5575.93896250606"
[1] "k= 17 0.972489561104916 6066.98821601118"
[1] "k= 19 0.970156297423221 6564.80508932371"
[1] "k= 21 0.968159141080858 7040.79195210361"
[1] "k= 23 0.963122972556195 7972.64791578296"
[1] "k= 25 0.9635670752405 8032.37822502446"
[1] "k= 27 0.963222969825831 8113.40101313052"
[1] "k= 29 0.961791341311107 8428.0823145298"
[1] "k= 31 0.961104760727688 8750.53111666039"
[1] "k= 33 0.960091948584439 9094.63946673805"
[1] "k= 35 0.958517368092966 9501.31780704318"
[1] "k= 37 0.956327483006558 9889.88697467776"
[1] "k= 39 0.954798489339322 10404.5673664586"
```

Hide

```
plot(1:20, cor_k, lwd=2, col='green', ylab="", yaxt='n')
par(new=TRUE)
```

Hide

```
plot(1:20, mse_k, lwd=2, col='purple', ylab="", labels=FALSE, yaxt='n')
```



It appears when k=3 we got the lowest MSE and the highest correlation, just to reiterate, let’s fit a knn regression with k=3 and see our results.

Hide

```
fit <- knnreg(train_scaled, train$price_x, k=3)
pred <- predict(fit, test_scaled)
cor_knn <- cor(pred, test$price_x)
mse_knn <- mean((pred-test$price_x)^2)
print(paste("cor=", cor_knn))
```

```
[1] "cor= 0.994586889110888"
```

Hide

```
print(paste("mse=", mse_knn))
```

```
[1] "mse= 1021.90929102559"
```

In so far, KNN regression appears to be heavily outperforming linear regression. Let’s see if decision tree regression can improve it further.

Decision Tree Regression

We will use the same columns used in the other two regressions.

Hide

```
set.seed(125) # reproducibility
i <- sample(1:nrow(data), nrow(data)*0.80, replace=FALSE)
train <- data[i,]
test <- data[-i,]
```

Hide

```
tree1 <- tree(price_x~.,data=train)
summary(tree1)
```

```


Regression tree:
tree(formula = price_x ~ ., data = train)
Variables actually used in tree construction:
[1] "bedrooms"  "bathrooms"
Number of terminal nodes:  3
Residual mean deviance:  72110 = 4.829e+09 / 66960
Distribution of residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
 -440.500   -56.150   -32.800     0.000     7.853 19840.000
```

Hide

```
pred <- predict(tree1,newdata=test)
cor_tree <- cor(pred, test$price_x)
mse_tree <- mean((pred-test$price_x)^2)
print(paste("cor=",cor_tree))
```

```
[1] "cor= 0.216114372156003"
```

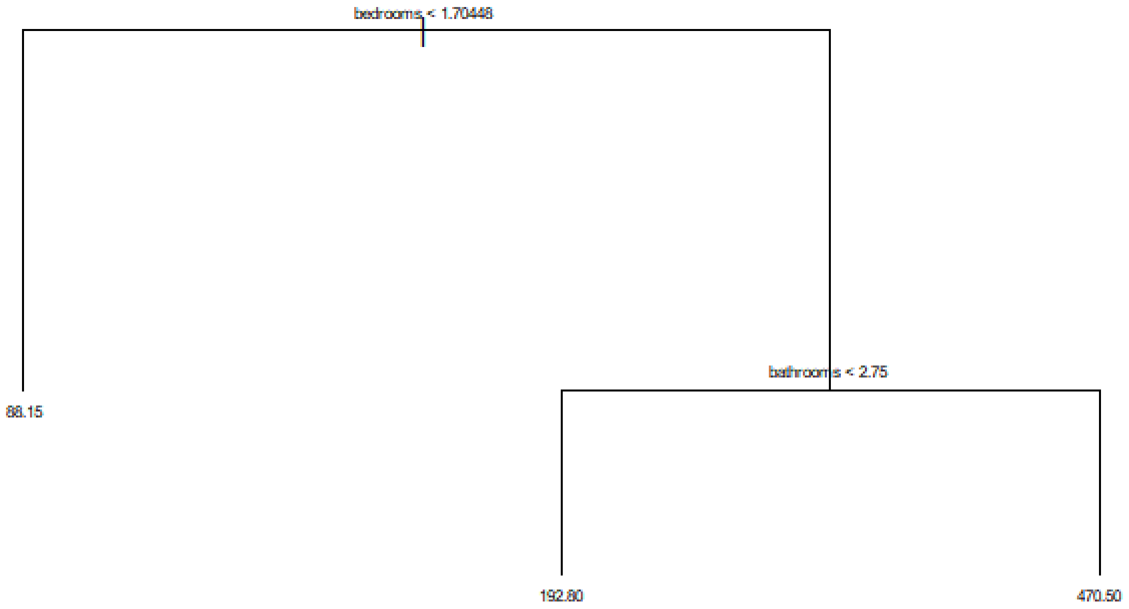
Hide

```
print(paste("mse=",mse_tree))
```

```
[1] "mse= 85121.8317102931"
```

Hide

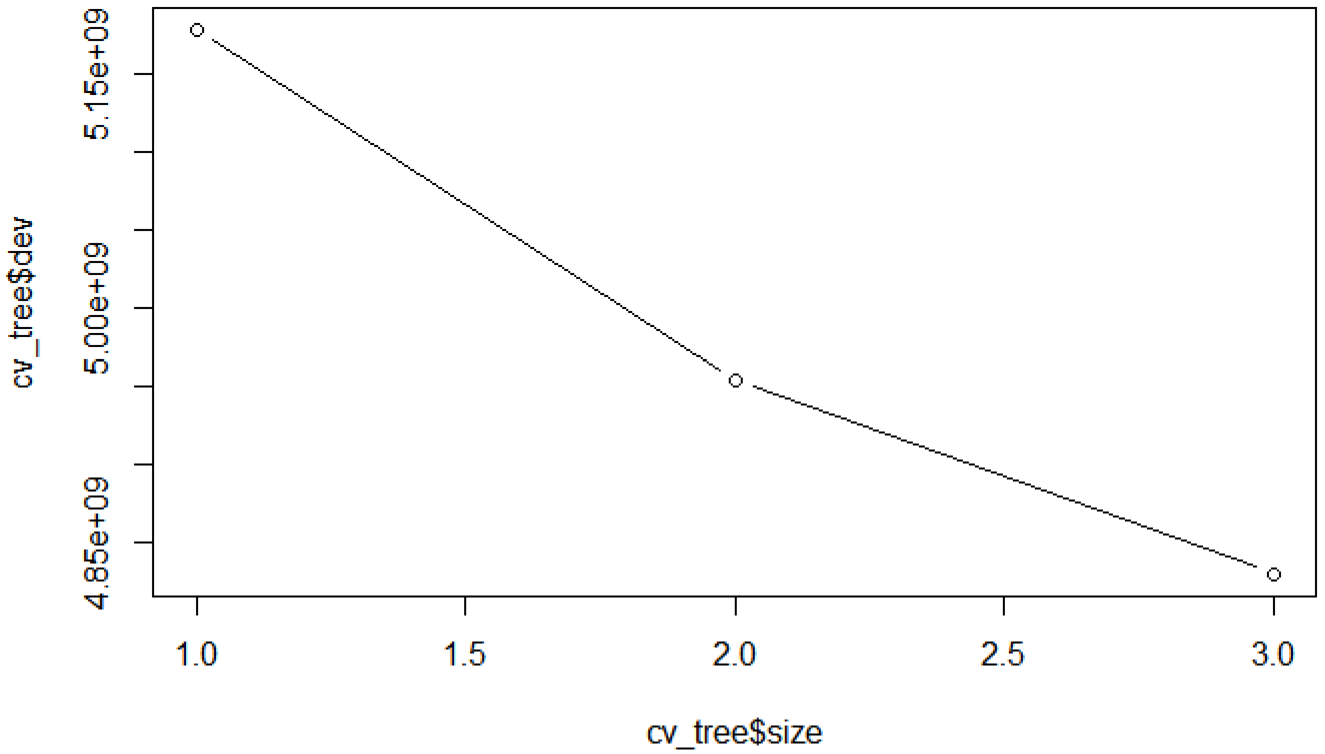
```
plot(tree1)
text(tree1, cex=0.5, pretty=0)
```



The tree only has 3 nodes, it’s deviance is quite high, let’s inspect cross validation.

Hide

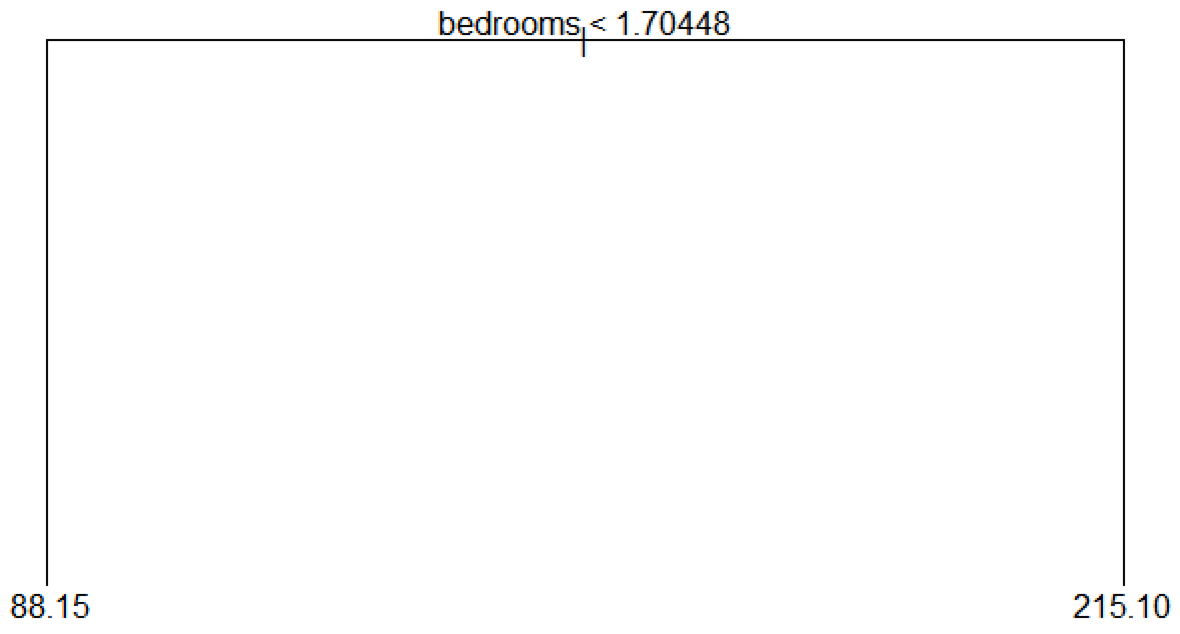
```
cv_tree <- cv.tree(tree1)
plot(cv_tree$size, cv_tree$dev, type='b')
```



The middle ground appears to be at 2 nodes, pruning the tree to 2 terminal nodes, then plotting

Hide

```
tree_pruned <- prune.tree(tree1, best=2)
plot(tree_pruned)
text(tree_pruned, pretty=0)
```



The correlation and mse for the original tree were the worst in so far. Measuring for the pruned tree, let’s see if it improves.

Hide

```
pred_pruned <- predict(tree_pruned, newdata=test)
cor_pruned <- cor(pred_pruned, test$price_x)
mse_pruned <- mean((pred_pruned-test$price_x)^2)
print(paste("corr=",cor_pruned))
```

```
[1] "corr= 0.184066304579861"
```

Hide

```
print(paste("mse=",mse_pruned))
```

```
[1] "mse= 86212.5684800797"
```

Nope, apparently not. The correlation went down and the mse raised. This was the worst result of all the regressions.

Analysis

Out of the three models explored in this notebook: linear regression, knn regression, and decision tree regression, the most performant was knn regression. There was apparent clustering as a result of data exploration. We could clearly see clustering relationships between the target and multiple predictors, showing a similar relationship. The goodness of fit for the linear regression was poor as there were many outliers present in the data. The knn clustering algorithm helped reduce the group size to just 3 clusters, while minimizing the mean absolute error and maximizing the correlation. The decision tree regression wasn't as helpful in this regard as a highly deviant tree of only three nodes was constructed, leaving the middle ground of two terminal nodes to be as deviant.