ECON573 Final Project

Nicholas Wong

November 19, 2023

Data cleaning/pre-processing

```r
set.seed(123)

# Mutate original price variable
airbnb <- airbnb |>
  mutate(price = exp(log_price),
         log_price = NULL)

# Drop NAs
airbnb <- na.omit(airbnb)

# host_response_rate to numeric for easier interpretation
airbnb <- airbnb |>
  mutate(host_response_rate = str_replace_all(host_response_rate, pattern = "%", replacement = "")) |>
  mutate_at(12, as.numeric)

# typecast to factors for fitting boosting
airbnb <- airbnb |>
  mutate(property_type = as.factor(property_type),
         room_type = as.factor(room_type),
         bed_type = as.factor(bed_type),
         cancellation_policy = as.factor(cancellation_policy),
         cleaning_fee = as.factor(cleaning_fee),
         city = as.factor(city),
         host_has_profile_pic = as.factor(host_has_profile_pic),
         host_identity_verified = as.factor(host_identity_verified),
         instant_bookable = as.factor(instant_bookable))


# 80/20 train/test split
training_indices <- sample(1:nrow(airbnb), .8*nrow(airbnb))

# Split data into train and test sets
train <- airbnb[training_indices, ]
test <- airbnb[-training_indices, ]

# relevel factors for train/test CV
totalData <- rbind(train, test)
for (f in 1:length(names(totalData))) {
  levels(train[, f]) <- levels(totalData[, f])
```

```
  levels(test[,f]) <- levels(totalData[, f])
}
```

Method 1 - Forward Selection

Here, we use a validation set approach due to the heavy computational expense of using stepwise methods with K-fold CV.

Fitting forward selection on training set

```
full = lm(price ~., data=train)
none = lm(price ~1, data = train)
MSE = (summary(full)$sigma)^2
forward_selection_mod <- step(none, scope = list(upper = full), scale = MSE, direction = 'forward', trac
```

```
## Start:  AIC=45000.59
## price ~ 1
##
##                          Df Sum of Sq       RSS    Cp
## + accommodates            1 243054866 456926061 16104
## + bedrooms                1 226055945 473924981 18125
## + bathrooms               1 178600170 521380757 23767
## + beds                    1 172420944 527559983 24502
## + room_type               2 140769981 559210946 28267
## + cancellation_policy     4  23467468 676513459 42218
## + property_type          31  14155524 685825403 43380
## + city                    5  12812083 687168844 43487
## + cleaning_fee            1   9870974 690109953 43829
## + bed_type                4   3316584 696664343 44614
## + longitude               1   2702227 697278699 44681
## + review_scores_rating    1   2498628 697482299 44706
## + host_since              1   2119253 697861673 44751
## + last_review             1   2012549 697968378 44763
## + number_of_reviews       1   1798098 698182828 44789
## + first_review            1   1408476 698572450 44835
## + instant_bookable        1   1192889 698788037 44861
## + host_identity_verified  1   1050314 698930613 44878
## + latitude                1    490904 699490023 44944
## + host_has_profile_pic    1    322235 699658691 44964
## <none>                                699980927 45001
## + host_response_rate      1     11171 699969756 45001
##
## Step:  AIC=16103.76
## price ~ accommodates
##
##                          Df Sum of Sq       RSS    Cp
## + bathrooms               1  41071612 415854449 11222
## + bedrooms                1  28997589 427928471 12658
## + room_type               2  25276770 431649290 13102
## + city                    5  18297736 438628324 13938
## + review_scores_rating    1   4603528 452322532 15558
## + property_type          31   4866912 452059149 15587
```

```
## + cancellation_policy     4   4180172 452745888 15615
## + instant_bookable        1   3556056 453370004 15683
## + host_since              1   3253074 453672986 15719
## + number_of_reviews       1   2105716 454820344 15855
## + first_review            1   1429663 455496398 15936
## + last_review             1   1200568 455725493 15963
## + host_response_rate      1    347145 456578915 16064
## + cleaning_fee            1    329674 456596387 16067
## + longitude               1    287070 456638991 16072
## + beds                    1    264802 456661258 16074
## + bed_type                4    299205 456626856 16076
## + latitude                1    105622 456820439 16093
## + host_identity_verified  1     88536 456837525 16095
## + host_has_profile_pic    1     66681 456859380 16098
## <none>                              456926061 16104
##
## Step:  AIC=11222.42
## price ~ accommodates + bathrooms
##
##                         Df Sum of Sq       RSS      Cp
## + room_type              2  38274516 377579933  6675.6
## + city                   5  19105706 396748743  8960.8
## + bedrooms               1  10855966 404998483  9933.7
## + property_type         31   8122383 407732066 10318.7
## + review_scores_rating   1   3946462 411907987 10755.2
## + cancellation_policy    4   3780607 412073841 10780.9
## + host_since             1   3040675 412813774 10862.9
## + instant_bookable       1   2773342 413081107 10894.7
## + first_review           1   1284405 414570044 11071.7
## + number_of_reviews      1   1067713 414786736 11097.5
## + latitude               1    970074 414884375 11109.1
## + cleaning_fee           1    660359 415194090 11145.9
## + beds                   1    625530 415228919 11150.0
## + host_response_rate     1    329979 415524470 11185.2
## + bed_type               4    297957 415556492 11195.0
## + last_review            1    225419 415629029 11197.6
## + host_identity_verified 1    106867 415747582 11211.7
## + host_has_profile_pic   1     98488 415755961 11212.7
## <none>                              415854449 11222.4
## + longitude              1     11069 415843380 11223.1
##
## Step:  AIC=6675.64
## price ~ accommodates + bathrooms + room_type
##
##                         Df Sum of Sq       RSS     Cp
## + city                   5  18533132 359046800 4482.1
## + bedrooms               1  14970549 362609383 4897.7
## + property_type         31   4888912 372691021 6156.4
## + review_scores_rating   1   2626321 374953612 6365.4
## + cancellation_policy    4   2537814 375042118 6381.9
## + instant_bookable       1   1644633 375935299 6482.1
## + host_since             1   1530663 376049270 6495.6
## + latitude               1   1509928 376070005 6498.1
## + number_of_reviews      1   1032691 376547241 6554.9
```

```
## + first_review           1    624513 376955420 6603.4
## + host_response_rate     1    306868 377273065 6641.2
## + last_review            1    149565 377430368 6659.9
## + longitude              1    135160 377444772 6661.6
## + host_has_profile_pic   1    121152 377458781 6663.2
## + beds                   1    114222 377465711 6664.1
## + bed_type               4    121440 377458493 6669.2
## + cleaning_fee           1     40075 377539857 6672.9
## <none>                              377579933 6675.6
## + host_identity_verified  1      4607 377575326 6677.1
##
## Step:  AIC=4482.08
## price ~ accommodates + bathrooms + room_type + city
##
##                          Df Sum of Sq       RSS     Cp
## + bedrooms                1  13271747 345775053 2906.1
## + longitude               1   9555284 349491516 3348.0
## + property_type          31   4846365 354200435 3967.9
## + last_review             1   4183079 354863722 3986.7
## + review_scores_rating    1   2888274 356158526 4140.7
## + cancellation_policy     4   2288417 356758383 4218.0
## + instant_bookable        1   1651028 357395772 4287.8
## + number_of_reviews       1   1565460 357481340 4297.9
## + host_since              1    958173 358088628 4370.2
## + first_review            1    627575 358419225 4409.5
## + host_response_rate      1    317628 358729172 4446.3
## + beds                    1    227941 358818860 4457.0
## + latitude                1    179557 358867244 4462.7
## + cleaning_fee            1    113801 358932999 4470.5
## + host_identity_verified  1     92268 358954532 4473.1
## + bed_type                4    142686 358904114 4473.1
## + host_has_profile_pic    1     82061 358964740 4474.3
## <none>                              359046800 4482.1
##
## Step:  AIC=2906.09
## price ~ accommodates + bathrooms + room_type + city + bedrooms
##
##                          Df Sum of Sq       RSS     Cp
## + longitude               1   9565425 336209628 1770.8
## + property_type          31   4600664 341174390 2421.1
## + last_review             1   3377604 342397450 2506.5
## + review_scores_rating    1   2460814 343314239 2615.5
## + cancellation_policy     4   2193117 343581936 2653.3
## + beds                    1   1849627 343925426 2688.2
## + instant_bookable        1   1105655 344669398 2776.6
## + number_of_reviews       1    995583 344779470 2789.7
## + host_since              1    654577 345120476 2830.3
## + first_review            1    450461 345324592 2854.5
## + host_response_rate      1    283131 345491922 2874.4
## + latitude                1    239523 345535531 2879.6
## + host_has_profile_pic    1    149989 345625064 2890.3
## + cleaning_fee            1    141547 345633507 2891.3
## + bed_type                4    145572 345629481 2896.8
## + host_identity_verified  1     77406 345697648 2898.9
```

```
## <none>                                345775053 2906.1
##
## Step:  AIC=1770.78
## price ~ accommodates + bathrooms + room_type + city + bedrooms +
##     longitude
##
##                          Df Sum of Sq       RSS     Cp
## + property_type          31   5346465 330863163 1197.1
## + last_review             1   3124623 333085005 1401.3
## + review_scores_rating    1   2266786 333942841 1503.3
## + cancellation_policy     4   2021315 334188313 1538.4
## + beds                    1   1788082 334421545 1560.2
## + number_of_reviews       1   1272934 334936694 1621.4
## + instant_bookable        1    830856 335378772 1674.0
## + host_since              1    322443 335887185 1734.4
## + cleaning_fee            1    243509 335966119 1743.8
## + host_response_rate      1    229142 335980486 1745.5
## + first_review            1    227665 335981963 1745.7
## + host_has_profile_pic    1    141947 336067681 1755.9
## + latitude                1     89981 336119647 1762.1
## + bed_type                4    131375 336078253 1763.2
## + host_identity_verified  1     37284 336172344 1768.3
## <none>                                336209628 1770.8
##
## Step:  AIC=1197.09
## price ~ accommodates + bathrooms + room_type + city + bedrooms +
##     longitude + property_type
##
##                          Df Sum of Sq       RSS      Cp
## + last_review             1   2998401 327864762  842.59
## + review_scores_rating    1   2064269 328798894  953.65
## + cancellation_policy     4   1735316 329127847  998.76
## + beds                    1   1360485 329502678 1037.33
## + number_of_reviews       1   1205001 329658162 1055.82
## + instant_bookable        1    741019 330122144 1110.99
## + host_since              1    287404 330575759 1164.92
## + cleaning_fee            1    272041 330591122 1166.75
## + host_response_rate      1    221872 330641291 1172.71
## + first_review            1    215049 330648114 1173.52
## + host_has_profile_pic    1    127873 330735290 1183.89
## + latitude                1    113298 330749865 1185.62
## + bed_type                4    108271 330754892 1192.22
## + host_identity_verified  1     30048 330833115 1195.52
## <none>                                330863163 1197.09
##
## Step:  AIC=842.59
## price ~ accommodates + bathrooms + room_type + city + bedrooms +
##     longitude + property_type + last_review
##
##                          Df Sum of Sq       RSS     Cp
## + review_scores_rating    1   2389393 325475369 560.49
## + cancellation_policy     4   1561723 326303039 664.90
## + beds                    1   1262404 326602358 694.49
## + number_of_reviews       1    619499 327245263 770.93
```

```
## + instant_bookable        1    331051 327533711 805.22
## + cleaning_fee            1    218288 327646474 818.63
## + host_has_profile_pic    1    123430 327741332 829.91
## + latitude                1    101195 327763567 832.55
## + host_since              1     63306 327801456 837.06
## + bed_type                4     83924 327780838 840.61
## + host_response_rate      1     16934 327847829 842.57
## <none>                              327864762 842.59
## + host_identity_verified  1     13020 327851742 843.04
## + first_review            1      4389 327860373 844.06
##
## Step:  AIC=560.49
## price ~ accommodates + bathrooms + room_type + city + bedrooms +
##     longitude + property_type + last_review + review_scores_rating
##
##                          Df Sum of Sq       RSS     Cp
## + cancellation_policy     4   1656102 323819267 371.58
## + beds                    1   1182126 324293244 421.94
## + number_of_reviews       1    592528 324882841 492.04
## + cleaning_fee            1    232932 325242437 534.80
## + instant_bookable        1    195772 325279597 539.21
## + host_has_profile_pic    1    129618 325345751 547.08
## + latitude                1    101286 325374083 550.45
## + host_response_rate      1     62039 325413330 555.11
## + host_since              1     34856 325440513 558.35
## + bed_type                4     73591 325401778 559.74
## <none>                              325475369 560.49
## + first_review            1     12017 325463352 561.06
## + host_identity_verified  1      1417 325473952 562.32
##
## Step:  AIC=371.58
## price ~ accommodates + bathrooms + room_type + city + bedrooms +
##     longitude + property_type + last_review + review_scores_rating +
##     cancellation_policy
##
##                          Df Sum of Sq       RSS     Cp
## + beds                    1   1144421 322674845 237.51
## + number_of_reviews       1    630139 323189128 298.66
## + cleaning_fee            1    296903 323522364 338.28
## + instant_bookable        1    204583 323614684 349.26
## + host_has_profile_pic    1    134706 323684561 357.57
## + latitude                1     97064 323722202 362.04
## + host_response_rate      1     66481 323752786 365.68
## + bed_type                4     75501 323743766 370.61
## + host_since              1     20928 323798338 371.10
## <none>                              323819267 371.58
## + first_review            1      2723 323816544 373.26
## + host_identity_verified  1       231 323819036 373.56
##
## Step:  AIC=237.51
## price ~ accommodates + bathrooms + room_type + city + bedrooms +
##     longitude + property_type + last_review + review_scores_rating +
##     cancellation_policy + beds
##
```

```
##                          Df Sum of Sq       RSS     Cp
## + number_of_reviews       1    597374 322077471 168.49
## + cleaning_fee            1    324535 322350311 200.93
## + instant_bookable        1    176462 322498383 218.53
## + host_has_profile_pic    1    132457 322542388 223.76
## + latitude                1     93253 322581592 228.43
## + host_response_rate      1     58351 322616494 232.58
## <none>                              322674845 237.51
## + host_since              1     13645 322661201 237.89
## + bed_type                4     56141 322618704 238.84
## + first_review            1      2905 322671941 239.17
## + host_identity_verified  1       254 322674591 239.48
##
## Step:  AIC=168.49
## price ~ accommodates + bathrooms + room_type + city + bedrooms +
##     longitude + property_type + last_review + review_scores_rating +
##     cancellation_policy + beds + number_of_reviews
##
##                          Df Sum of Sq       RSS     Cp
## + first_review            1    366341 321711130 126.93
## + cleaning_fee            1    349860 321727611 128.89
## + instant_bookable        1    158051 321919420 151.69
## + host_has_profile_pic    1    126322 321951149 155.47
## + host_since              1    101110 321976361 158.46
## + latitude                1     83309 321994162 160.58
## + host_response_rate      1     39911 322037560 165.74
## <none>                              322077471 168.49
## + host_identity_verified  1     13048 322064423 168.94
## + bed_type                4     55291 322022180 169.91
##
## Step:  AIC=126.93
## price ~ accommodates + bathrooms + room_type + city + bedrooms +
##     longitude + property_type + last_review + review_scores_rating +
##     cancellation_policy + beds + number_of_reviews + first_review
##
##                          Df Sum of Sq       RSS      Cp
## + cleaning_fee            1    362250 321348880  85.858
## + host_has_profile_pic    1    129135 321581995 113.575
## + latitude                1     82096 321629034 119.168
## + instant_bookable        1     77846 321633285 119.674
## + host_response_rate      1     34325 321676806 124.848
## <none>                              321711130 126.929
## + bed_type                4     53330 321657800 128.589
## + host_since              1      1682 321709448 128.729
## + host_identity_verified  1      1412 321709718 128.762
##
## Step:  AIC=85.86
## price ~ accommodates + bathrooms + room_type + city + bedrooms +
##     longitude + property_type + last_review + review_scores_rating +
##     cancellation_policy + beds + number_of_reviews + first_review +
##     cleaning_fee
##
##                          Df Sum of Sq       RSS     Cp
## + host_has_profile_pic    1    128559 321220321 72.573
```

```
## + latitude               1     76716 321272164 78.737
## + instant_bookable       1     73207 321275673 79.154
## + host_response_rate     1     24270 321324610 84.973
## <none>                               321348880 85.858
## + host_identity_verified  1      6148 321342732 87.127
## + bed_type               4     55436 321293445 87.267
## + host_since             1      3836 321345044 87.402
##
## Step:  AIC=72.57
## price ~ accommodates + bathrooms + room_type + city + bedrooms +
##     longitude + property_type + last_review + review_scores_rating +
##     cancellation_policy + beds + number_of_reviews + first_review +
##     cleaning_fee + host_has_profile_pic
##
##                          Df Sum of Sq       RSS     Cp
## + latitude               1     75907 321144414 65.548
## + instant_bookable       1     73899 321146422 65.787
## + host_response_rate     1     22895 321197426 71.851
## <none>                               321220321 72.573
## + host_identity_verified  1      9814 321210507 73.406
## + bed_type               4     55980 321164341 73.917
## + host_since             1      4699 321215622 74.014
##
## Step:  AIC=65.55
## price ~ accommodates + bathrooms + room_type + city + bedrooms +
##     longitude + property_type + last_review + review_scores_rating +
##     cancellation_policy + beds + number_of_reviews + first_review +
##     cleaning_fee + host_has_profile_pic + latitude
##
##                          Df Sum of Sq       RSS     Cp
## + instant_bookable       1     77029 321067385 58.389
## + host_response_rate     1     23412 321121002 64.764
## <none>                               321144414 65.548
## + host_identity_verified  1     10248 321134166 66.329
## + host_since             1      5328 321139086 66.914
## + bed_type               4     54224 321090190 67.100
##
## Step:  AIC=58.39
## price ~ accommodates + bathrooms + room_type + city + bedrooms +
##     longitude + property_type + last_review + review_scores_rating +
##     cancellation_policy + beds + number_of_reviews + first_review +
##     cleaning_fee + host_has_profile_pic + latitude + instant_bookable
##
##                          Df Sum of Sq       RSS     Cp
## + host_response_rate     1     17144 321050241 58.351
## <none>                               321067385 58.389
## + host_identity_verified  1      6761 321060624 59.585
## + bed_type               4     53338 321014047 60.047
## + host_since             1      2720 321064665 60.066
##
## Step:  AIC=58.35
## price ~ accommodates + bathrooms + room_type + city + bedrooms +
##     longitude + property_type + last_review + review_scores_rating +
##     cancellation_policy + beds + number_of_reviews + first_review +
```

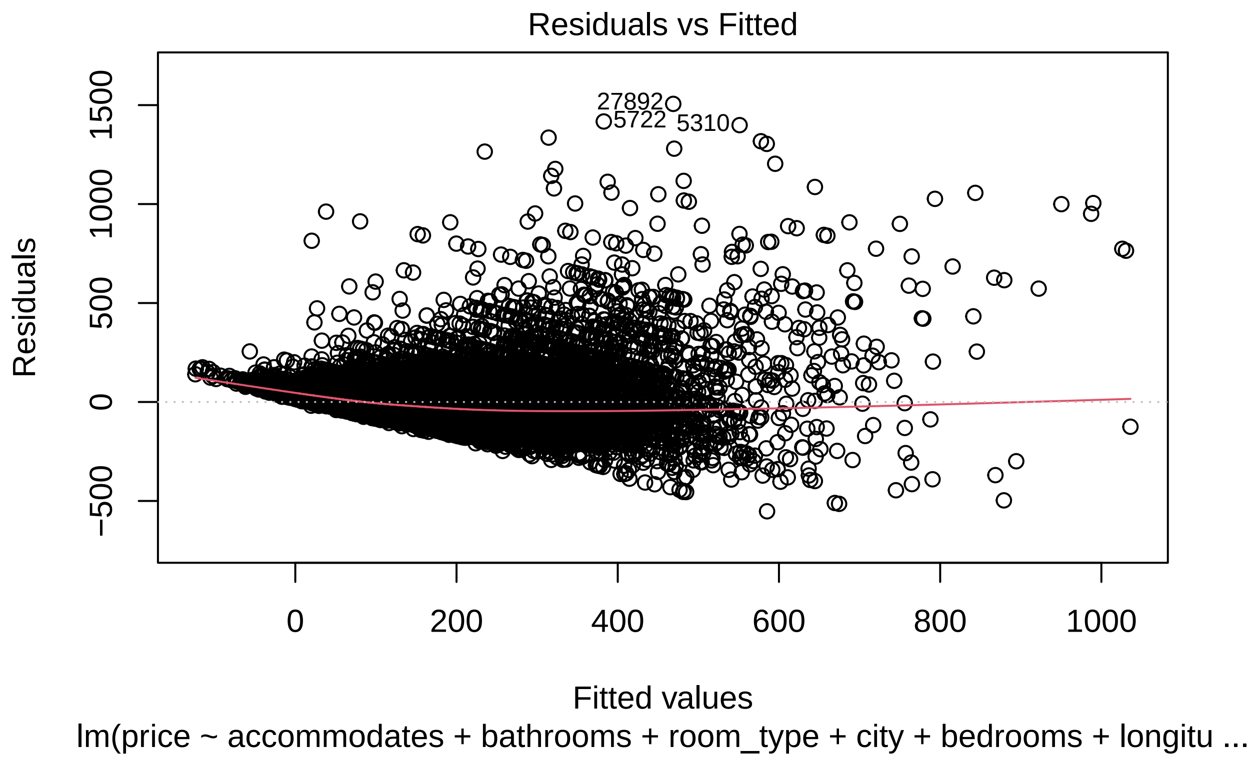
```
##     cleaning_fee + host_has_profile_pic + latitude + instant_bookable +
##     host_response_rate
##
##                          Df Sum of Sq       RSS     Cp
## <none>                                321050241 58.351
## + host_identity_verified  1      7382 321042859 59.473
## + bed_type                4     53479 320996763 59.992
## + host_since              1      2786 321047455 60.020
```

```r
summary(forward_selection_mod)
```

```
##
## Call:
## lm(formula = price ~ accommodates + bathrooms + room_type + city +
##     bedrooms + longitude + property_type + last_review + review_scores_rating +
##     cancellation_policy + beds + number_of_reviews + first_review +
##     cleaning_fee + host_has_profile_pic + latitude + instant_bookable +
##     host_response_rate, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -552.24  -41.89   -5.95   28.34 1506.14
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      -1.234e+04  4.592e+02 -26.869  < 2e-16 ***
## accommodates                      1.490e+01  4.369e-01  34.106  < 2e-16 ***
## bathrooms                         6.099e+01  1.044e+00  58.399  < 2e-16 ***
## room_typePrivate room            -6.910e+01  1.185e+00 -58.321  < 2e-16 ***
## room_typeShared room             -1.025e+02  3.135e+00 -32.705  < 2e-16 ***
## cityChicago                      -2.914e+03  8.581e+01 -33.956  < 2e-16 ***
## cityDC                           -9.822e+02  3.974e+01 -24.715  < 2e-16 ***
## cityLA                           -8.022e+03  2.535e+02 -31.647  < 2e-16 ***
## cityNYC                          -4.594e+02  1.909e+01 -24.062  < 2e-16 ***
## citySF                           -8.736e+03  2.686e+02 -32.525  < 2e-16 ***
## bedrooms                          3.395e+01  8.997e-01  37.732  < 2e-16 ***
## longitude                        -1.729e+02  5.160e+00 -33.503  < 2e-16 ***
## property_typeBed & Breakfast      7.704e+00  5.692e+00   1.353 0.175913
## property_typeBoat                 5.072e+01  1.512e+01   3.355 0.000795 ***
## property_typeBoutique hotel       4.014e+00  1.742e+01   0.230 0.817746
## property_typeBungalow             1.916e+00  6.239e+00   0.307 0.758781
## property_typeCabin                5.996e-01  1.456e+01   0.041 0.967163
## property_typeCamper/RV           -2.001e+01  1.303e+01  -1.536 0.124589
## property_typeCastle               5.764e+01  2.768e+01   2.083 0.037298 *
## property_typeCave                 2.514e+01  9.174e+01   0.274 0.784041
## property_typeChalet              -2.537e+00  4.105e+01  -0.062 0.950716
## property_typeCondominium          1.476e+01  2.554e+00   5.779 7.55e-09 ***
## property_typeDorm                -4.847e+01  1.043e+01  -4.645 3.41e-06 ***
## property_typeEarth House          3.561e+01  6.486e+01   0.549 0.583022
## property_typeGuest suite         -5.165e+00  1.070e+01  -0.483 0.629359
## property_typeGuesthouse          -2.670e+00  5.183e+00  -0.515 0.606448
## property_typeHostel              -7.098e+01  1.384e+01  -5.129 2.92e-07 ***
## property_typeHouse                2.717e+00  1.280e+00   2.123 0.033798 *
## property_typeHut                  7.380e-01  4.104e+01   0.018 0.985651
```

```
## property_typeIn-law                   -2.826e+01  1.256e+01  -2.250 0.024476 *
## property_typeIsland                    1.128e+02  9.186e+01   1.227 0.219651
## property_typeLoft                      4.118e+01  3.520e+00  11.698  < 2e-16 ***
## property_typeOther                     1.194e+01  5.269e+00   2.266 0.023445 *
## property_typeServiced apartment        6.928e+01  2.453e+01   2.825 0.004735 **
## property_typeTent                     -4.037e+01  2.769e+01  -1.458 0.144883
## property_typeTimeshare                 1.045e+02  1.876e+01   5.573 2.52e-08 ***
## property_typeTipi                      8.761e+01  5.299e+01   1.653 0.098275 .
## property_typeTownhouse                -5.949e+00  3.083e+00  -1.930 0.053606 .
## property_typeTrain                     4.782e+01  6.489e+01   0.737 0.461204
## property_typeTreehouse                 2.675e+02  5.299e+01   5.048 4.48e-07 ***
## property_typeVacation home             3.519e+01  4.589e+01   0.767 0.443090
## property_typeVilla                     1.295e+02  9.321e+00  13.899  < 2e-16 ***
## property_typeYurt                     -2.543e+01  3.748e+01  -0.679 0.497393
## last_review                           -4.232e-02  3.418e-03 -12.381  < 2e-16 ***
## review_scores_rating                   1.117e+00  6.639e-02  16.818  < 2e-16 ***
## cancellation_policymoderate            3.855e-01  1.422e+00   0.271 0.786285
## cancellation_policystrict              6.835e+00  1.342e+00   5.095 3.51e-07 ***
## cancellation_policysuper_strict_30  4.516e+01  1.139e+01   3.965 7.37e-05 ***
## cancellation_policysuper_strict_60  4.179e+02  3.483e+01  12.001  < 2e-16 ***
## beds                                  -7.712e+00  6.751e-01 -11.424  < 2e-16 ***
## number_of_reviews                     -1.409e-01  1.383e-02 -10.182  < 2e-16 ***
## first_review                          -7.260e-03  1.221e-03  -5.945 2.79e-09 ***
## cleaning_feeTRUE                      -8.296e+00  1.299e+00  -6.388 1.70e-10 ***
## host_has_profile_picTRUE              -4.615e+01  1.187e+01  -3.889 0.000101 ***
## latitude                               2.106e+01  6.855e+00   3.072 0.002129 **
## instant_bookableTRUE                  -3.157e+00  1.088e+00  -2.901 0.003728 **
## host_response_rate                    -5.246e-02  3.674e-02  -1.428 0.153393
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 91.71 on 38171 degrees of freedom
## Multiple R-squared:  0.5413, Adjusted R-squared:  0.5407
## F-statistic: 804.5 on 56 and 38171 DF,  p-value: < 2.2e-16
```

```r
plot(forward_selection_mod, 1)
```



Get predictions and residuals for forward selection, compute MSE

```
test_forwardselection <- test |> add_predictions(forward_selection_mod, var = "forward_pred")
test_forwardselection <- test_forwardselection |> add_residuals(forward_selection_mod, var = "forward_r

# Args: vector of residuals
# Return: RMSE
RMSE_func <- function(resid){
  return(sqrt(mean(resid^2)))
}

(forward_selection_RMSE <- RMSE_func(test_forwardselection$forward_resid))
```

```
## [1] 98.32442
```

```
tidy(forward_selection_mod) |>
  filter(p.value < .05)
```

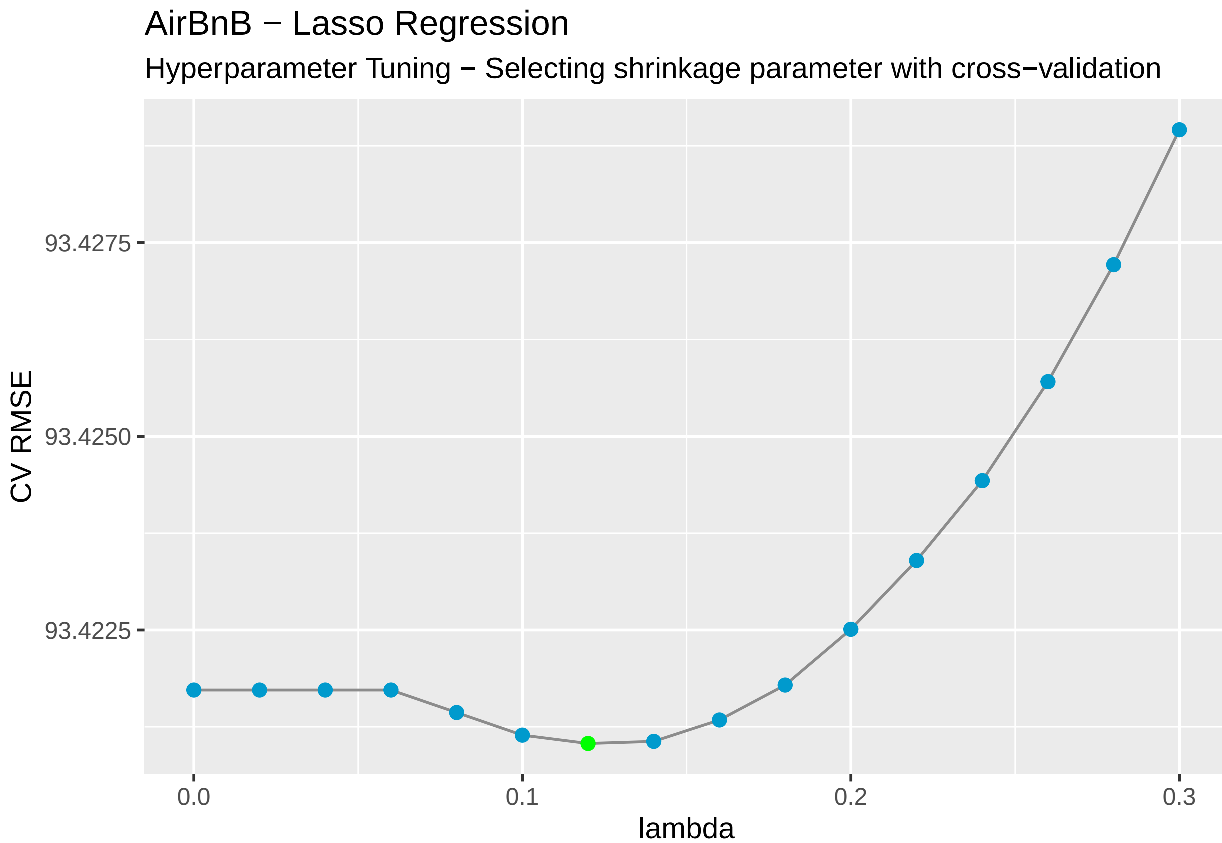
```
## # A tibble: 37 x 5
##    term                  estimate std.error statistic   p.value
##    <chr>                    <dbl>     <dbl>     <dbl>     <dbl>
##  1 (Intercept)           -12339.    459.        -26.9 1.48e-157
##  2 accommodates              14.9     0.437      34.1 3.60e-251
##  3 bathrooms                 61.0     1.04       58.4 0
##  4 room_typePrivate room    -69.1     1.18      -58.3 0
```

```
##   5 room_typeShared room    -103.      3.14      -32.7 2.13e-231
##   6 cityChicago           -2914.     85.8       -34.0 5.06e-249
##   7 cityDC                 -982.     39.7       -24.7 8.39e-134
##   8 cityLA                -8022.    253.        -31.6 5.43e-217
##   9 cityNYC                -459.     19.1       -24.1 5.51e-127
##  10 citySF                -8736.    269.        -32.5 6.49e-229
## # i 27 more rows
```

Method 2 - LASSO

Here, we select shrinkage parameter λ for LASSO through repeated 5-fold CV. We test a range of 16 different λ values in (0, 0.3), in equally spaced increments of 0.02.

```r
ctrl <- trainControl(method = "repeatedcv", number = 5, repeats = 10, verboseIter = F)

set.seed(1)

model_lasso <- train(price ~ .,
                   data = train,
                   method = "glmnet",
                   preProcess = c("center", "scale"),
                     metric = "RMSE",
                     maximize = F,
                     trControl = ctrl,
                     tuneGrid = expand.grid(alpha = 1, # lasso specification
                                            lambda = seq(0, 0.3, 0.02)))

model_lasso$results |>
  rename(CV_RMSE = RMSE) |>
  mutate(min_CV_RMSE = as.numeric(lambda == model_lasso$bestTune$lambda)) |>
  ggplot(aes(x = lambda, y = CV_RMSE)) +
  geom_line(col = "grey55") +
  geom_point(size = 2, aes(col = factor(min_CV_RMSE))) +
  scale_color_manual(values = c("deepskyblue3", "green")) +
  theme(legend.position = "none") +
  labs(title = "AirBnB - Lasso Regression",
       subtitle = "Hyperparameter Tuning - Selecting shrinkage parameter with cross-validation",
       y = "CV RMSE")
```



Optimal shrinkage (λ):

```
model_lasso$bestTune$lambda
```

```
## [1] 0.12
```

CV RMSE:

```
(lasso_cv <- min(model_lasso$results$RMSE) |>  round(4))
```

```
## [1] 93.421
```

Test RMSE:

```
(lasso_test_RMSE <- sqrt(mean((predict(model_lasso, test) - test$price)^2)) |>  round(4))
```

```
## [1] 100.0746
```

Predictors in final fitted LASSO model:

```
tibble(names = model_lasso$coefnames) |> kable()
```

names
property_typeBed & Breakfast
property_typeBoat
property_typeBoutique hotel
property_typeBungalow
property_typeCabin
property_typeCamper/RV
property_typeCastle
property_typeCave
property_typeChalet
property_typeCondominium
property_typeDorm
property_typeEarth House
property_typeGuest suite
property_typeGuesthouse
property_typeHostel
property_typeHouse
property_typeHut
property_typeIn-law
property_typeIsland
property_typeLoft
property_typeOther
property_typeServiced apartment
property_typeTent
property_typeTimeshare
property_typeTipi
property_typeTownhouse
property_typeTrain
property_typeTreehouse
property_typeVacation home
property_typeVilla
property_typeYurt
room_typePrivate room
room_typeShared room
accommodates
bathrooms
bed_typeCouch
bed_typeFuton
bed_typePull-out Sofa
bed_typeReal Bed
cancellation_policymoderate
cancellation_policystrict
cancellation_policysuper_strict_30
cancellation_policysuper_strict_60
cleaning_feeTRUE
cityChicago
cityDC
cityLA
cityNYC
citySF
first_review
host_has_profile_picTRUE
host_identity_verifiedTRUE

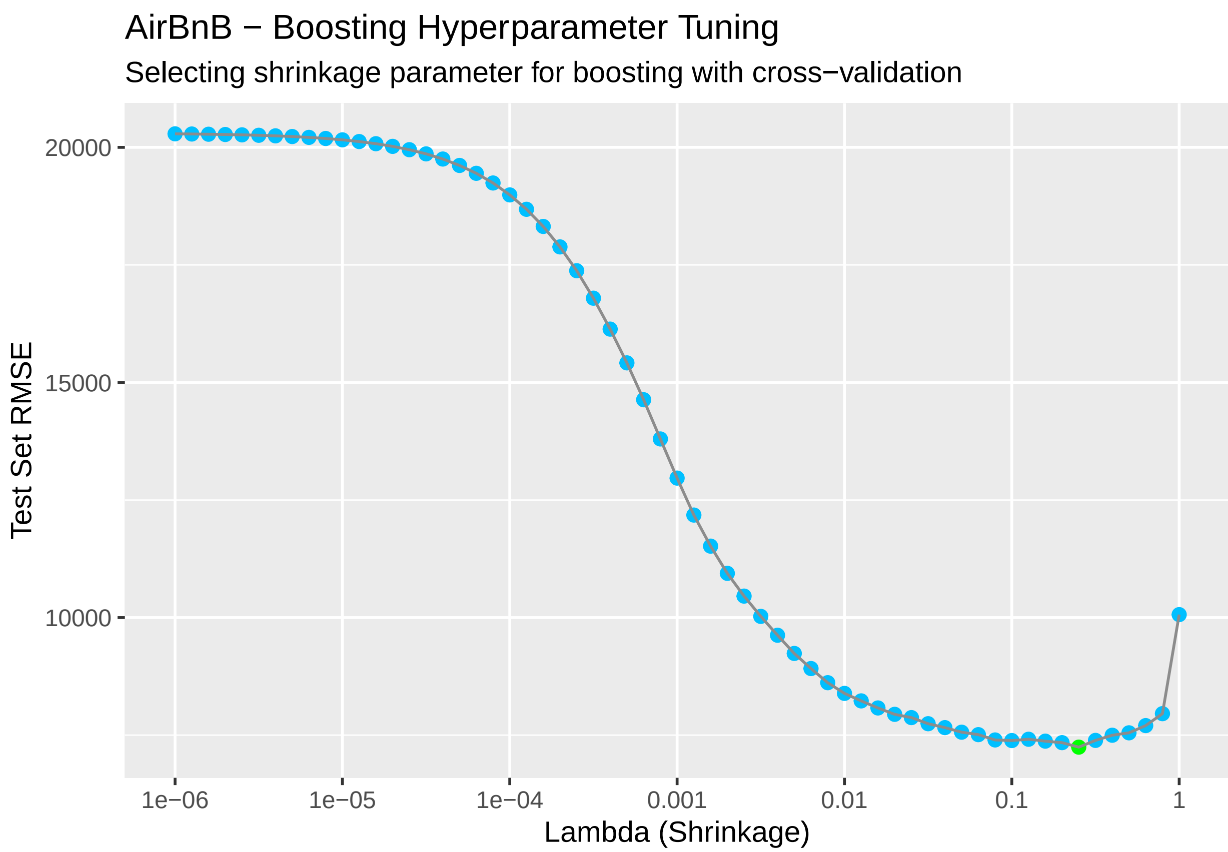
names
host_response_rate
host_since
instant_bookableTRUE
last_review
latitude
longitude
number_of_reviews
review_scores_rating
bedrooms
beds

Method 3 - Boosting

Validation set approach for gbm

```
lambda_seq <- 10^seq(-6, 0, 0.1)

set.seed(123)

train_MSE <- c()
test_MSE <- c()


for (i in 1:length(lambda_seq)) {
  boost_TEMP <- gbm(price ~ . -first_review -host_since -last_review,
                    data = train,
                    distribution = "gaussian",
                    n.trees = 1000,
                    interaction.depth = 2,
                    shrinkage = lambda_seq[i])

  train_MSE[i] <- mean((predict(boost_TEMP, train, n.trees = 1000) - train$price)^2)

  test_MSE[i] <- mean((predict(boost_TEMP, test, n.trees = 1000) - test$price)^2)
}

df <- data.frame(lambda = lambda_seq, test_MSE) |>
  mutate(min_MSE = as.numeric(test_MSE == min(test_MSE)))

df |>
  ggplot(aes(x = lambda, y = test_MSE)) +
  geom_point(size = 2, aes(col = factor(min_MSE))) +
  geom_line(col = "grey55") +
  scale_color_manual(values = c("deepskyblue", "green")) +
  theme(legend.position = "none") +
  scale_x_continuous(trans = 'log10', breaks = 10^seq(-6, 0), labels = 10^seq(-6, 0), minor_breaks = NU
  labs(x = "Lambda (Shrinkage)",
       y = "Test MSE") +
  labs(title = "AirBnB - Boosting Hyperparameter Tuning",
       subtitle = "Selecting shrinkage parameter for boosting with cross-validation",
```

```
       y = "Test Set RMSE")
```




```
(boosting_RMSE <- sqrt(df$test_MSE[which(df$min_MSE == 1)]))
```

```
## [1] 85.11441
```

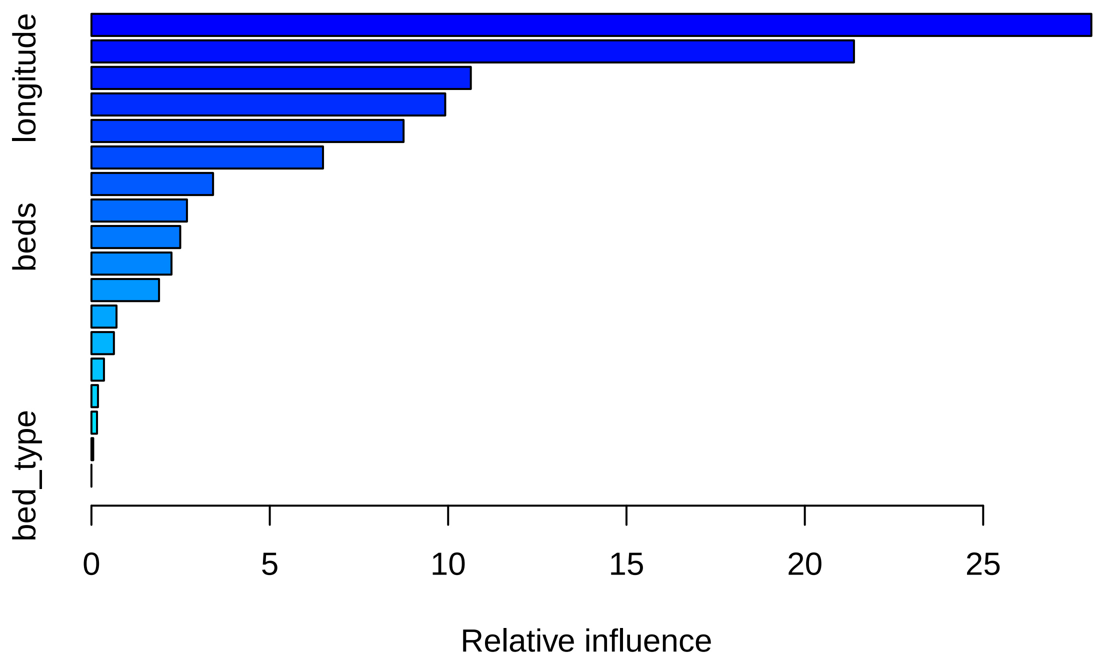
```
(boosting_lambda <- df$lambda[which(df$min_MSE == 1)])
```

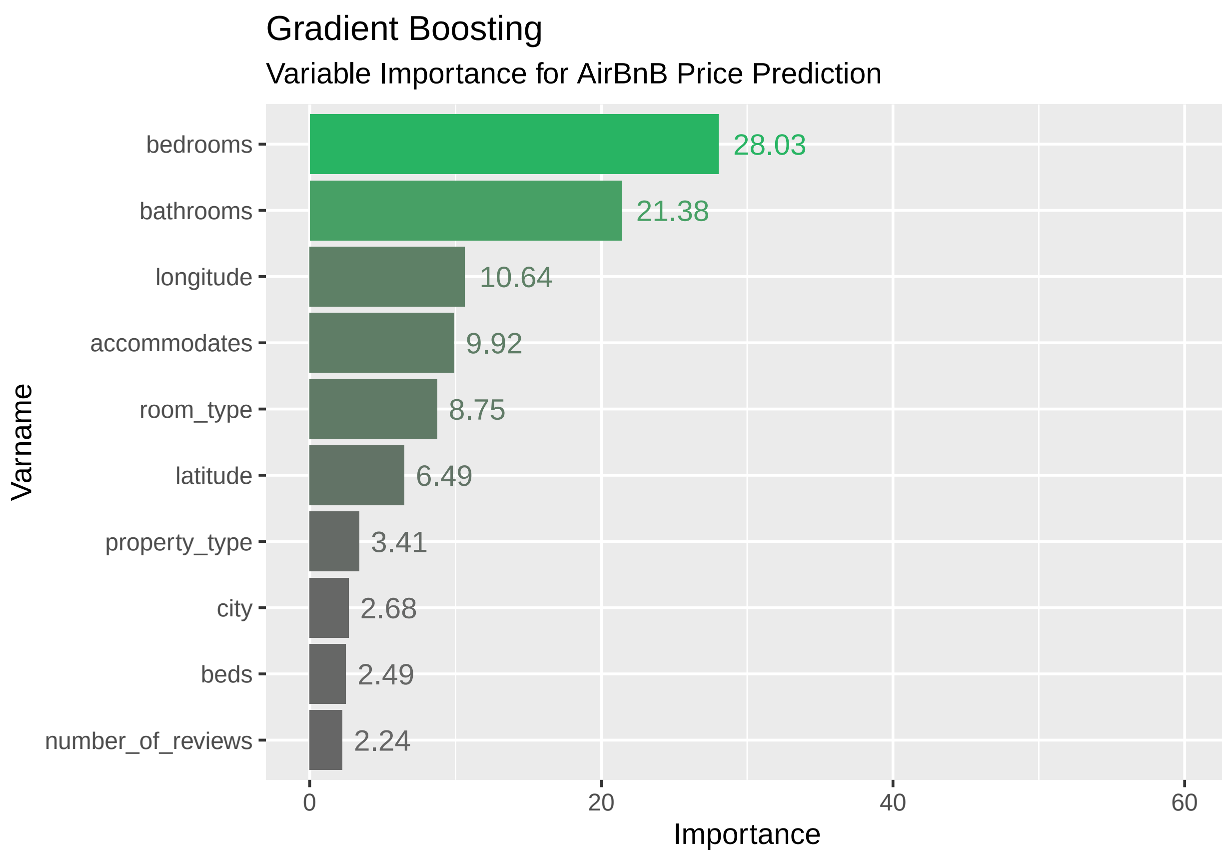
```
## [1] 0.2511886
```

Plot with optimal lambda using validation set approach

```
boost_TEMP <- gbm(price ~ . -first_review -host_since -last_review,
                    data = train,
                    distribution = "gaussian",
                    n.trees = 1000,
                    interaction.depth = 2,
                    shrinkage = boosting_lambda)

# ggplot version:
summary(boost_TEMP)[1:10,] |>
  rename("Importance" = "rel.inf") |>
  ggplot(aes(x = fct_reorder(var, Importance), y = Importance, fill = Importance)) +
```

```
  geom_bar(stat = "identity") +
  geom_text(aes(label = round(Importance, 2), col = Importance), hjust = -0.2) +
  scale_y_continuous(limits = c(0, 60)) +
  scale_fill_gradient(low = "grey40", high = "#28B463") +
  scale_color_gradient(low = "grey40", high = "#28B463") +
  coord_flip() +
  theme(legend.position = "none") +
  labs(title = "Gradient Boosting",
       subtitle = "Variable Importance for AirBnB Price Prediction",
       x = "Varname")
```

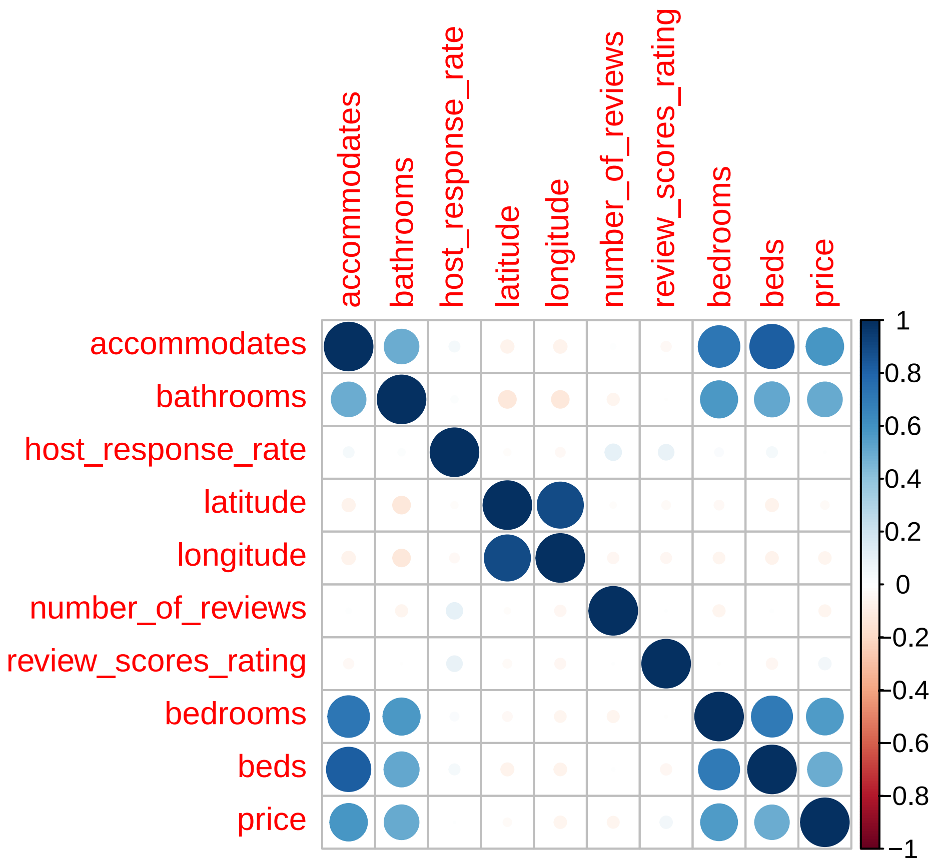




Method 4 - Polynomial

First, we perform an exploratory data analysis (EDA) to find the variable for polynomial fit

```
airbnb_numeric <- airbnb[, sapply(airbnb, is.numeric)]
corrplot::corrplot(cor(airbnb_numeric))
```

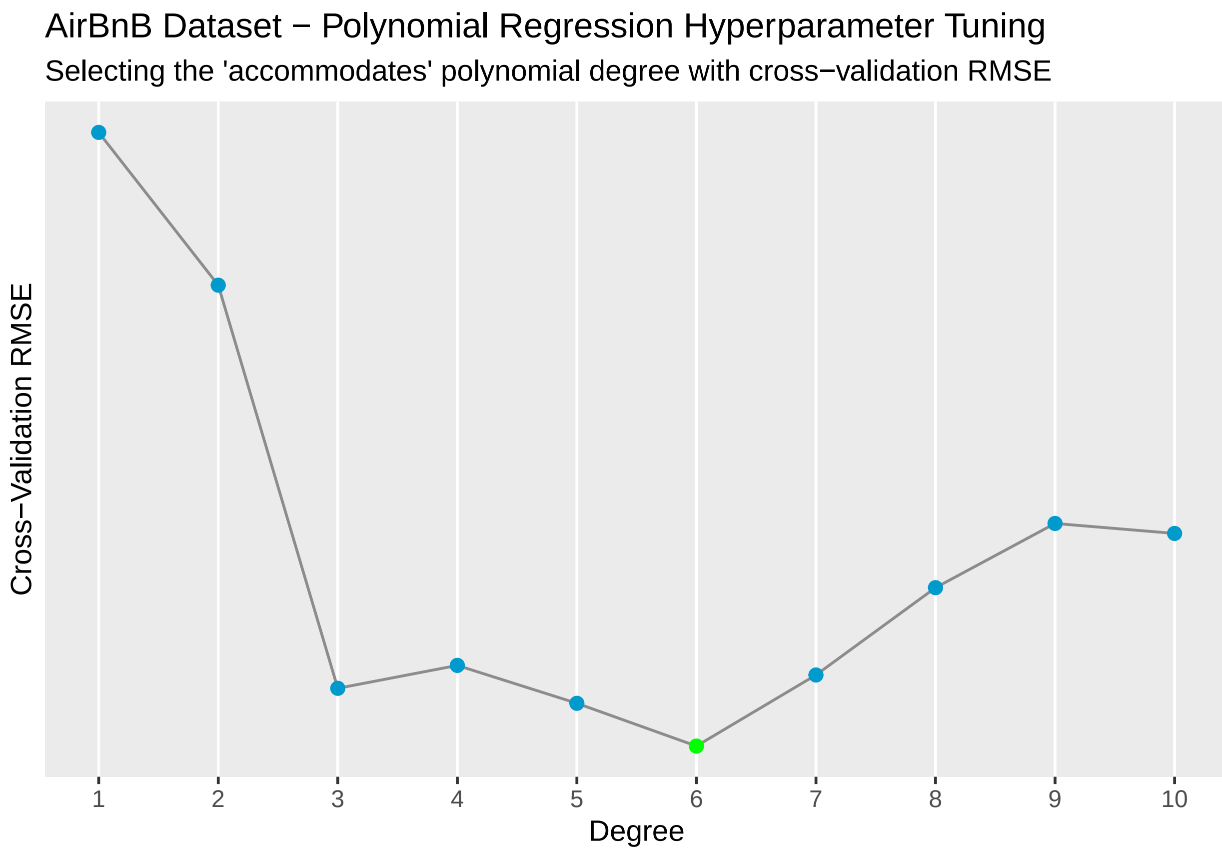


Highest correlation with "accommodates". Use accommodates for polynomial fit.

Here, we opt for K-fold CV to choose the optimal degree for the polynomial. We perform 10-fold CV with 5 repeats.

```
ctrl <- trainControl(method = "repeatedcv", number = 10, repeats = 5)

CV_RMSE <- c()

set.seed(159)

for (i in 1:10) {
  model_temp <- train(y = train$price,
                      x = poly(train$accommodates, i, raw = T, simple = T),
                      method = "lm",
                      metric = "RMSE",
                      trControl = ctrl)
  CV_RMSE[i] <- model_temp$results$RMSE
}

data.frame(degree = 1:10, CV_RMSE = CV_RMSE) |>
  mutate(min_CV_RMSE = as.numeric(min(CV_RMSE) == CV_RMSE)) |>
  ggplot(aes(x = degree, y = CV_RMSE)) +
  geom_line(col = "grey55") +
  geom_point(size = 2, aes(col = factor(min_CV_RMSE))) +
  scale_x_continuous(breaks = seq(1, 10), minor_breaks = NULL) +
  scale_y_continuous(breaks = seq(0, 0.03, 0.002)) +
```

```
scale_color_manual(values = c("deepskyblue3", "green")) +
theme(legend.position = "none") +
labs(title = "AirBnB Dataset - Polynomial Regression Hyperparameter Tuning",
     subtitle = "Selecting the 'accommodates' polynomial degree with cross-validation RMSE",
     x = "Degree",
     y = "Cross-Validation RMSE")
```



We find that polynomial degree 6 minimizes RMSE

```
# store minimum polynomial RMSE for reference
(min_poly_RMSE_raw <- min(CV_RMSE))
```

```
## [1] 108.9932
```

Now, we use a validation set approach to test on unseen test data

```
polymod <- lm(price ~ poly(accommodates, 6, raw = T), data = train)

test_poly <- test |> add_predictions(polymod, var = "poly_pred")
test_poly <- test_poly |> add_residuals(polymod, var = "poly_resid")

(poly_RMSE <- RMSE_func(test_poly$poly_resid))
```

```
## [1] 115.8819
```

```r
summary(polymod)
```

```
##
## Call:
## lm(formula = price ~ poly(accommodates, 6, raw = T), data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -494.68  -47.08  -16.43   27.40 1614.78
##
## Coefficients:
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      2.127e+01  9.088e+00   2.341   0.0193 *
## poly(accommodates, 6, raw = T)1  5.979e+01  1.317e+01   4.539 5.67e-06 ***
## poly(accommodates, 6, raw = T)2 -1.615e+01  6.926e+00  -2.332   0.0197 *
## poly(accommodates, 6, raw = T)3  3.915e+00  1.685e+00   2.323   0.0202 *
## poly(accommodates, 6, raw = T)4 -4.089e-01  2.030e-01  -2.014   0.0440 *
## poly(accommodates, 6, raw = T)5  1.919e-02  1.169e-02   1.642   0.1006
## poly(accommodates, 6, raw = T)6 -3.383e-04  2.556e-04  -1.323   0.1858
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 109 on 38221 degrees of freedom
## Multiple R-squared:  0.3508, Adjusted R-squared:  0.3507
## F-statistic:  3442 on 6 and 38221 DF,  p-value: < 2.2e-16
```

```r
#tidy(polymod)
```

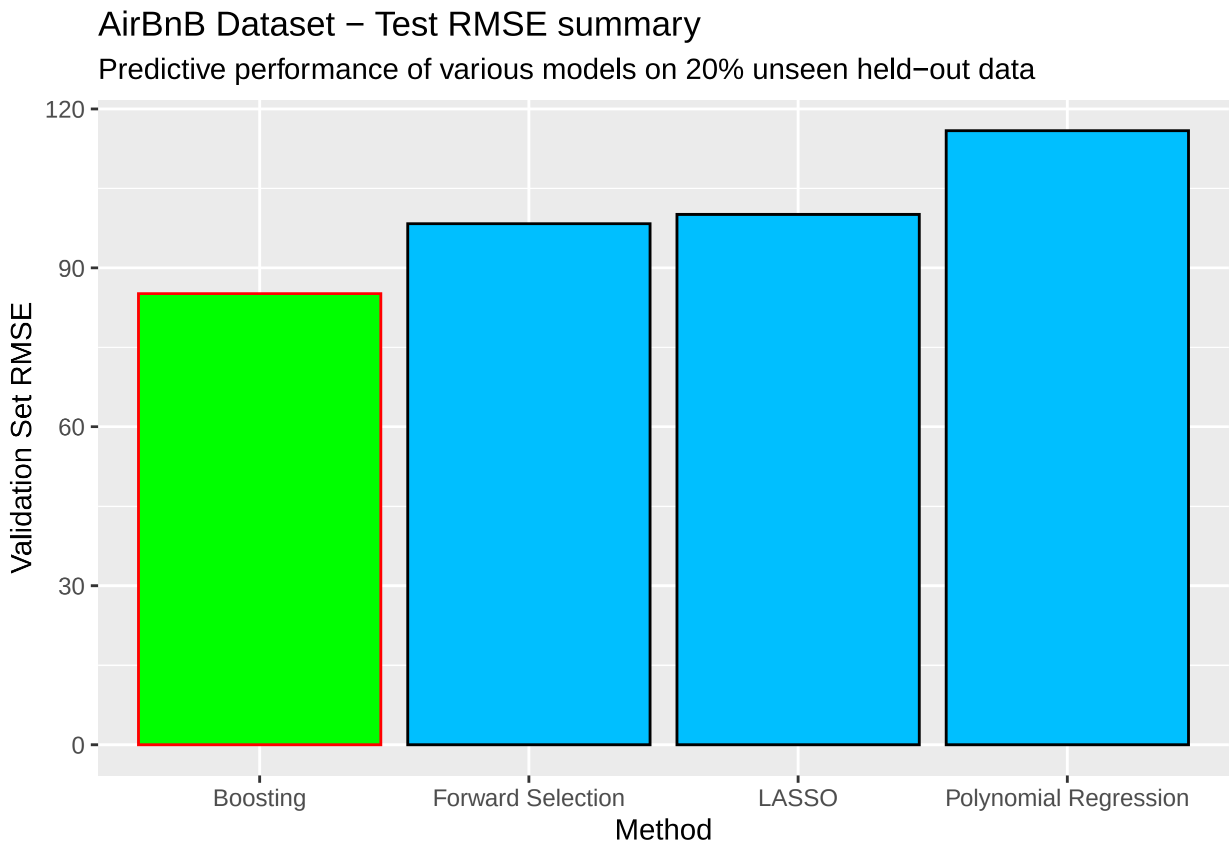
Summary of test error for methods

```r
(RMSE_summary <- tibble(Method = c("Forward Selection", "LASSO", "Polynomial Regression", "Boosting"), R
```

```
## # A tibble: 4 x 2
##   Method                 RMSE
##   <chr>                 <dbl>
## 1 Forward Selection      98.3
## 2 LASSO                 100.
## 3 Polynomial Regression 116.
## 4 Boosting               85.1
```

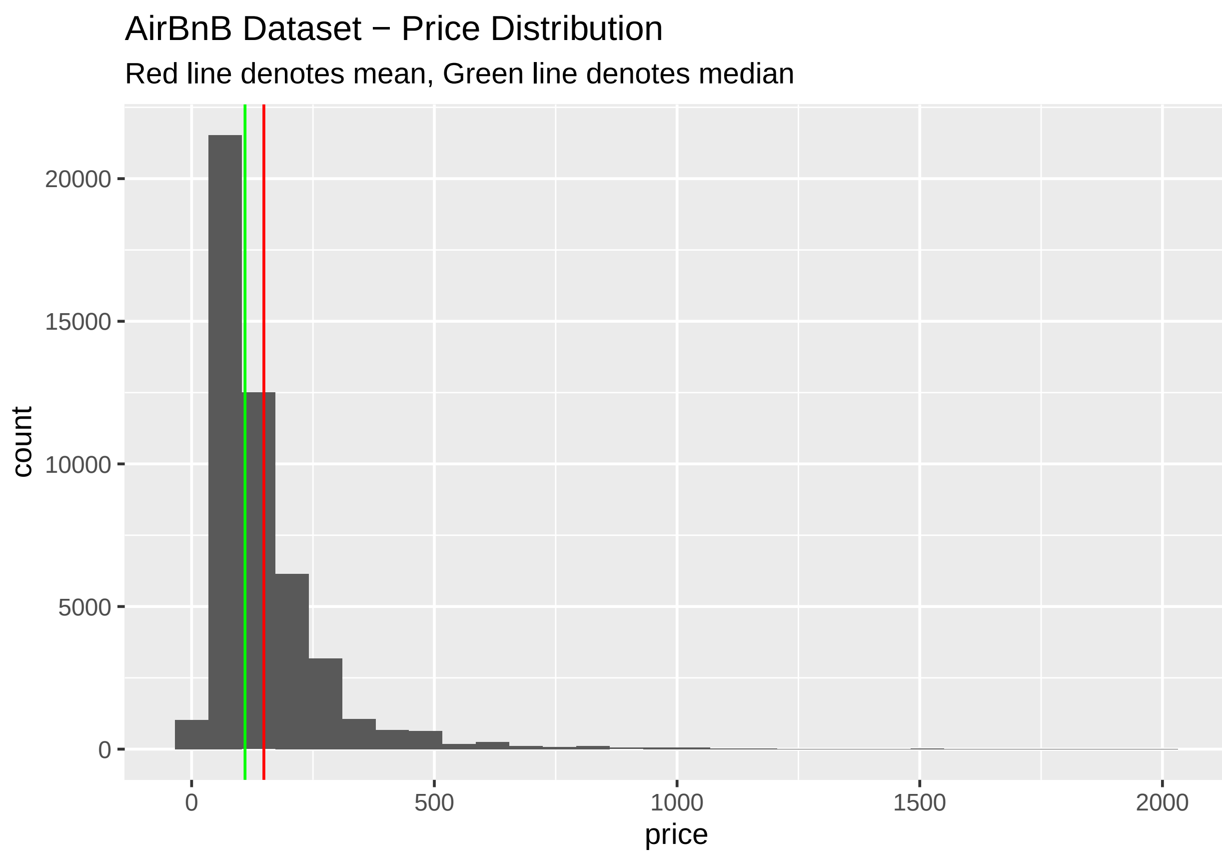
```r
RMSE_summary |>
  mutate(min_RMSE = as.numeric(min(RMSE) == RMSE)) |>
  ggplot(aes(x = Method, y = RMSE)) +
  geom_col(aes(fill = factor(min_RMSE), color = factor(min_RMSE))) +
  scale_fill_manual(values = c("deepskyblue", "green")) +
  scale_color_manual(values = c("black", "red")) +
  theme(legend.position = "none") +
  labs(title = "AirBnB Dataset - Test RMSE summary",
       subtitle = "Predictive performance of various models on 20% unseen held-out data",
```

```
       x = "Method",
       y = "Validation Set RMSE")
```

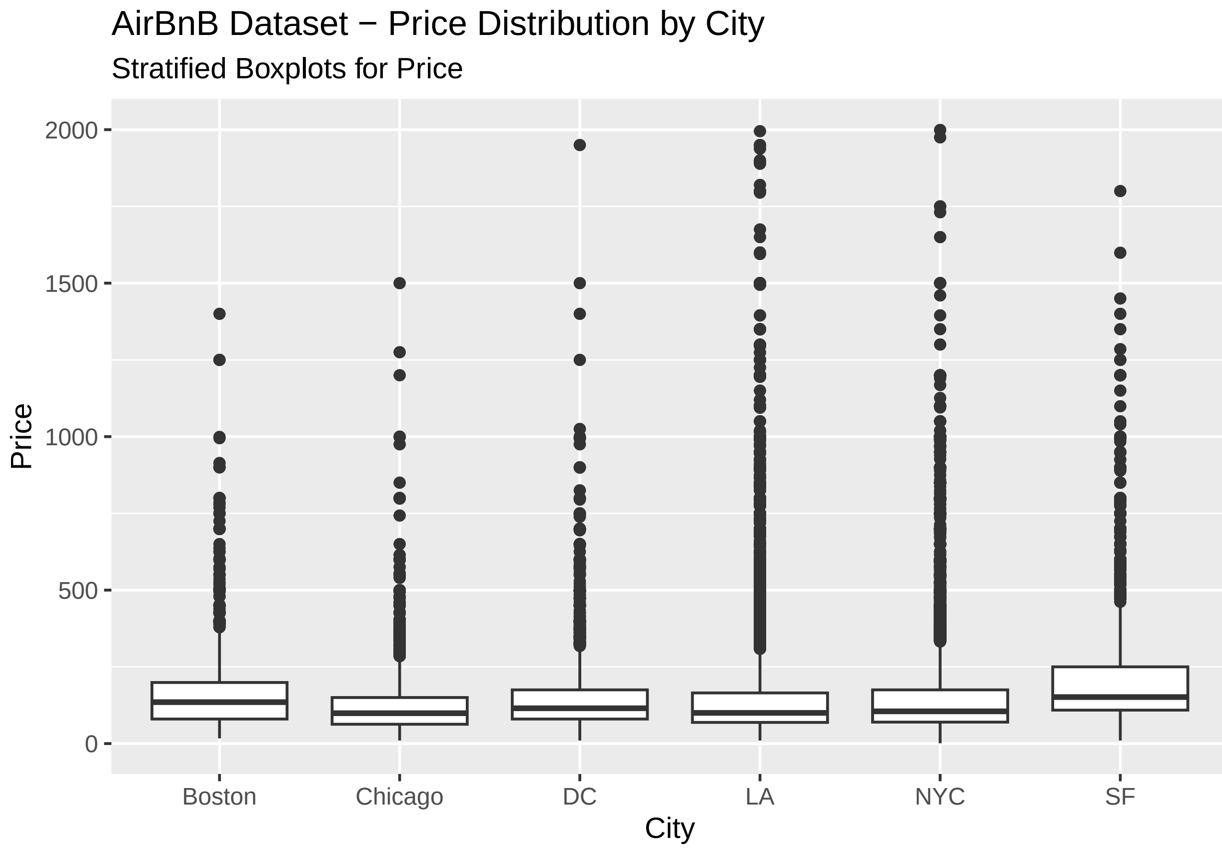


##Descriptive stats##

```
airbnb |>
  ggplot(aes(x = price)) +
  geom_histogram(bins = 30) +
  geom_vline(xintercept = mean(airbnb$price), color = "red") +
  geom_vline(xintercept = median(airbnb$price), color = "green") +
  labs(title = "AirBnB Dataset - Price Distribution",
       subtitle = "Red line denotes mean, Green line denotes median") +
  theme(legend.position = "bottom")
```




```r
airbnb |>
  ggplot(aes(x = city, y = price)) +
  geom_boxplot() +
  xlab("City") +
  ylab("Price") +
  labs(title = "AirBnB Dataset - Price Distribution by City",
       subtitle = "Stratified Boxplots for Price")
```



AirBnB Dataset − Price Distribution by City
Stratified Boxplots for Price
Price
2000
1500
1000
500
0
Boston
Chicago
DC
LA
NYC
SF
City