GW | Business

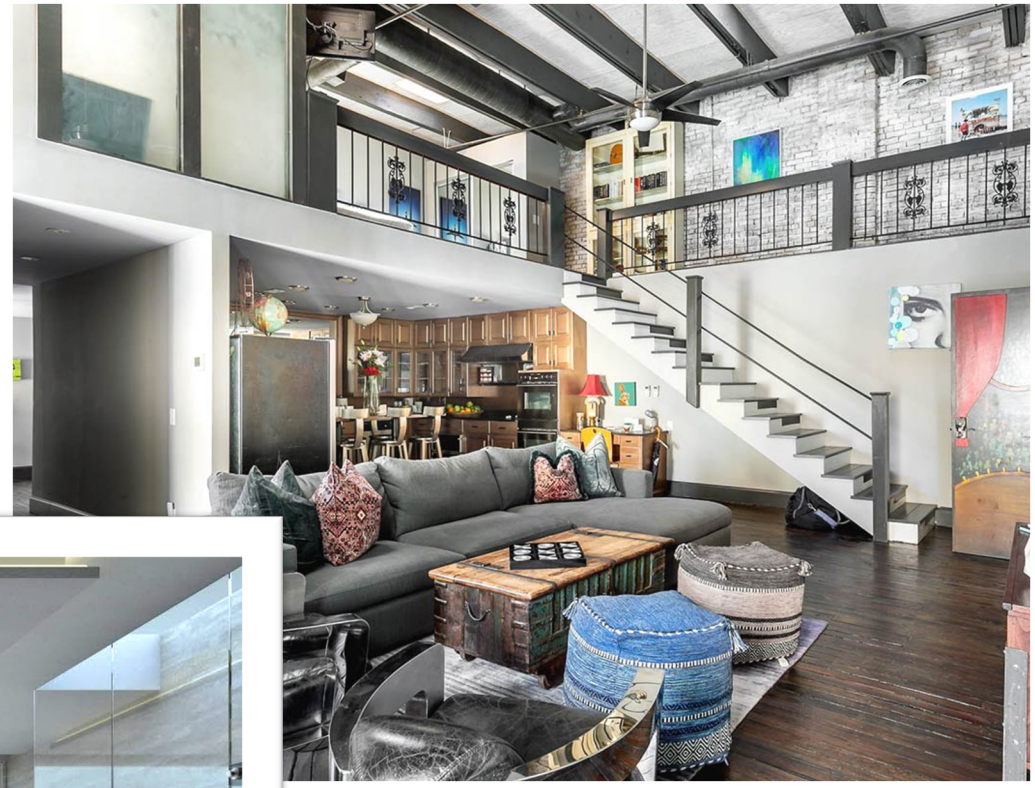
Digital analysis of USA housing price and type

Data Mining Group 8: *Shengqi Zhou ,Yichen Li, Fengshu Xu , Yangzhuopeng Yi*

Apirl 14, 2020

- Introduction
- Data description and preprocessing
- Methods
- Results
- Conclusion and remark

Introduction

- 



Introduction


```
     price                     type           sqfeet             beds
 Min.   :0.000e+00   apartment   :186097   Min.   :      0   Min.   :   0.000
 1st Qu.:8.190e+02   house       : 22219   1st Qu.:    750   1st Qu.:   1.000
 Median :1.059e+03   townhouse   : 12869   Median :    950   Median :   2.000
 Mean   :1.351e+04   condo       :  4711   Mean   :   1105   Mean   :   1.928
 3rd Qu.:1.464e+03   duplex      :  4490   3rd Qu.:   1154   3rd Qu.:   2.000
 Max.   :2.768e+09   manufactured:  3820   Max.   :8388607   Max.   :1100.000
                     (Other)     :  1764
     baths          cats_allowed     dogs_allowed    smoking_allowed wheelchair_access
 Min.   : 0.000   Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
 1st Qu.: 1.000   1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:0.000   1st Qu.:0.0000
 Median : 1.000   Median :1.0000   Median :1.0000   Median :1.000   Median :0.0000
 Mean   : 1.479   Mean   :0.7793   Mean   :0.7532   Mean   :0.643   Mean   :0.1042
 3rd Qu.: 2.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:0.0000
 Max.   :75.000   Max.   :1.0000   Max.   :1.0000   Max.   :1.000   Max.   :1.0000

 electric_vehicle_charge comes_furnished              laundry_options
 Min.   :0.00000         Min.   :0.00000                     :     0
 1st Qu.:0.00000         1st Qu.:0.00000   laundry in bldg   : 31409
 Median :0.00000         Median :0.00000   laundry on site   : 44419
 Mean   :0.01755         Mean   :0.05822   no laundry on site:  3355
 3rd Qu.:0.00000         3rd Qu.:0.00000   w/d hookups       : 54485
 Max.   :1.00000         Max.   :1.00000   w/d in unit       :102302

          parking_options        lat              long             state
 off-street parking:125105   Min.   :-43.53   Min.   :-163.89   ca     : 24175
 attached garage    : 38670   1st Qu.: 33.96   1st Qu.:-105.07   tx     : 15542
 carport            : 38478   Median : 38.59   Median : -89.40   fl     : 15232
 detached garage    : 16356   Mean   : 37.89   Mean   : -94.22   mi     :  9834
 street parking     : 15362   3rd Qu.: 41.74   3rd Qu.: -81.57   oh     :  9246
 no parking         :  1857   Max.   : 64.99   Max.   : 172.63   nc     :  8886
 (Other)            :   142                                      (Other):153055
```

- 22 variables,384,977 observations
- Numerical: rent per month, total square footage, latitude, and longitude
- Categorical: number of beds, number of bathrooms, house region, house type, states, cats allowed, dogs allowed, smoking allowed, wheelchair access allowed, electric vehicle charger, comes with furniture.

- Introduction
- **Data description and preprocessing**
- Methods
- Results
- Conclusion and remark

Data Preprocessing

- Removal of unwanted variables
- Handling missing data
- Managing unwanted outliers

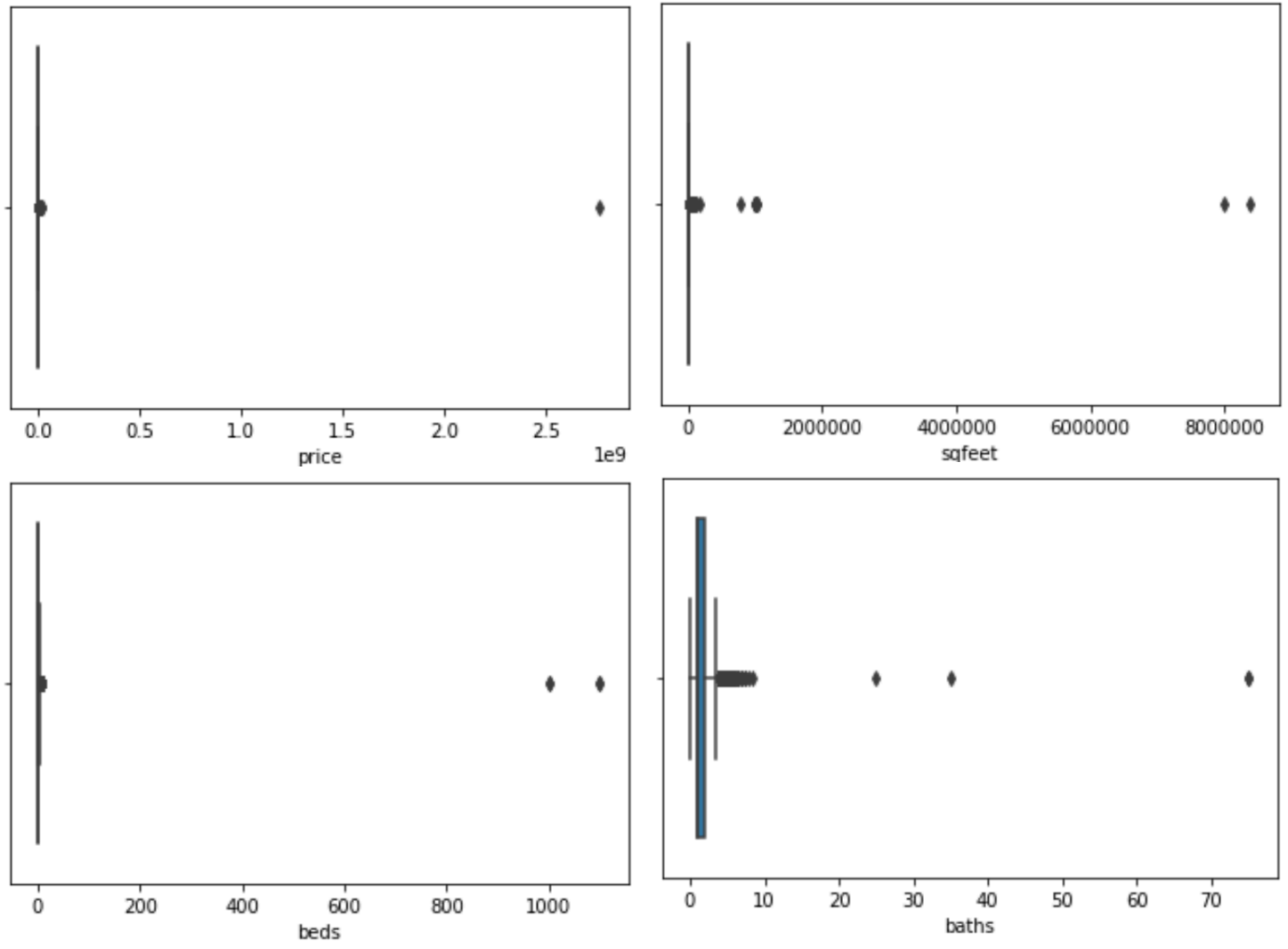
https://www.kaggle.com/austinreese/usa-housing-listings

Data Preprocessing

✓ Null Value

```
id                        False
url                       False
region                    False
region_url                False
price                     False
type                      False
sqfeet                    False
beds                      False
baths                     False
cats_allowed              False
dogs_allowed              False
smoking_allowed           False
wheelchair_access         False
electric_vehicle_charge   False
comes_furnished           False
laundry_options            True
parking_options            True
image_url                 False
description                True
lat                        True
long                       True
state                     False
dtype: bool
```

✓ Outliers



Data Preprocessing

Linear regression with outliers

```
Call:
lm(formula = price ~ sqfeet, data = house)

Residuals:
       Min         1Q     Median         3Q        Max
    -26027      -8020       -7789      -7430 2768298423

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 8.823e+03  7.203e+03   1.225    0.221
sqfeet      2.140e-03  3.755e-01   0.006    0.995

Residual standard error: 4462000 on 384975 degrees of freedom
Multiple R-squared:  8.437e-11, Adjusted R-squared:  -2.597e-06
F-statistic: 3.248e-05 on 1 and 384975 DF,  p-value: 0.9955
```

Linear regression without outliers

```
Call:
lm(formula = price ~ sqfeet, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-6448.3  -346.5  -137.4   203.7 12632.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.462e+02  3.154e+00   204.9   <2e-16 ***
sqfeet      5.775e-01  2.882e-03   200.4   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 594.1 on 233782 degrees of freedom
Multiple R-squared:  0.1466,    Adjusted R-squared:  0.1466
F-statistic: 4.015e+04 on 1 and 233782 DF,  p-value: < 2.2e-16
```

Data Preprocessing

- Change data types of few variables from numerical to categorical

- Unwanted columns: description, image_url

```
id                          int64
url                        object
region                     object
region_url                 object
price                       int64
type                       object
sqfeet                      int64
beds                        int64
baths                     float64
cats_allowed                int64
dogs_allowed                int64
smoking_allowed             int64
wheelchair_access           int64
electric_vehicle_charge     int64
comes_furnished             int64
laundry_options            object
parking_options            object
image_url                  object
description                object
lat                       float64
long                      float64
state                      object
dtype: object
```

USA housing Listing

- Introduction
- Data description and preprocessing
- Methods
- Results
- Conclusion and remark

Methods

- Least Square? Ridge Regression? LASSO?
- Variable Pre-Selection? Forward Stepwise Selection?
- Linear Regression? Support Vector Machines ? KNN?
- LDA & QDA? KNN classification?

  Support Vector Machines (SVM)?

Methods

- Least Square? Ridge Regression? LASSO?
- Variable Pre-Selection? Forward Stepwise Selection?
- Linear Regression? Support Vector Machines ? KNN
- LDA & QDA? KNN classification?

  Support Vector Machines (SVM)

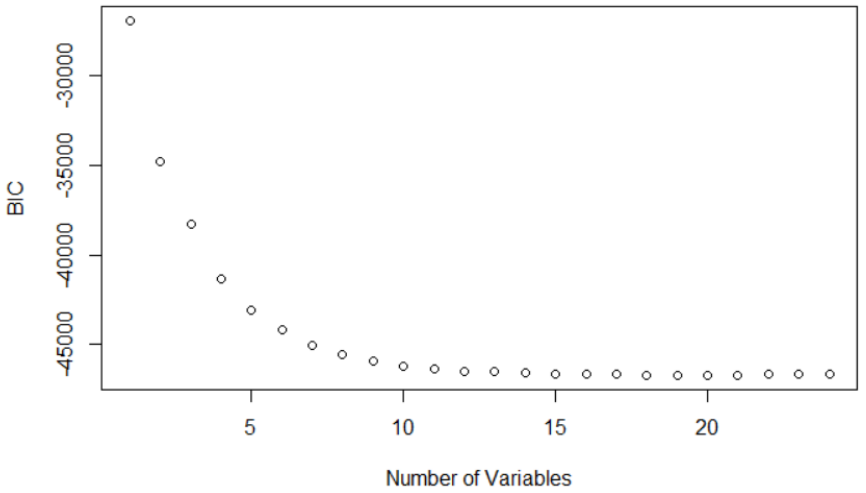
USA housing Listing

- Introduction
- Data description and preprocessing
- Methods
- **Results**
- Conclusion and remark

Results


```
Call:
lm(formula = price ~ sqfeet + type + +baths + dogs_allowed +
    electric_vehicle_charge + beds + laundry_options + parking_options +
    cats_allowed + smoking_allowed + comes_furnished + sqfeet *
    beds + sqfeet * baths, data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-6699.6  -283.9   -83.5   174.2 12840.3

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                           1.102e+03  8.037e+00 137.079  < 2e-16 ***
sqfeet                                2.254e-01  7.442e-03  30.284  < 2e-16 ***
typeassisted living                   4.034e+02  3.810e+02   1.059   0.2896
typecondo                             2.243e+02  8.227e+00  27.257  < 2e-16 ***
typecottage/cabin                     1.707e+02  2.246e+01   7.597 3.04e-14 ***
typeduplex                           -8.325e+01  8.440e+00  -9.863  < 2e-16 ***
typeflat                              1.830e+02  2.533e+01   7.226 5.00e-13 ***
typehouse                            -1.139e+02  4.932e+00 -23.097  < 2e-16 ***
typein-law                            2.537e+02  4.504e+01   5.633 1.78e-08 ***
typeland                             -5.782e+02  2.409e+02  -2.400   0.0164 *
typeloft                              9.021e+01  2.301e+01   3.920 8.85e-05 ***
typemanufactured                     -2.056e+02  9.187e+00 -22.379  < 2e-16 ***
typetownhouse                        -7.201e+01  5.149e+00 -13.984  < 2e-16 ***
baths                                 3.764e+01  5.141e+00   7.322 2.46e-13 ***
dogs_allowed                         -4.588e+01  4.832e+00  -9.495  < 2e-16 ***
electric_vehicle_charge               4.553e+02  8.633e+00  52.743  < 2e-16 ***
beds                                 -4.255e+01  3.486e+00 -12.204  < 2e-16 ***
laundry_optionslaundry on site        1.219e+01  4.048e+00   3.012   0.0026 **
laundry_optionsno laundry on site    -5.581e+01  9.990e+00  -5.586 2.33e-08 ***
laundry_optionsw/d hookups           -9.386e+01  4.123e+00 -22.765  < 2e-16 ***
laundry_optionsw/d in unit            2.254e+02  3.736e+00  60.347  < 2e-16 ***
parking_optionscarport               -1.605e+02  4.200e+00 -38.218  < 2e-16 ***
parking_optionsdetached garage       -1.440e+02  5.143e+00 -28.004  < 2e-16 ***
parking_optionsno parking            -6.772e+01  1.300e+01  -5.209 1.90e-07 ***
parking_optionsoff-street parking    -3.197e+02  3.576e+00 -89.392  < 2e-16 ***
parking_optionsstreet parking        -2.216e+02  5.395e+00 -41.081  < 2e-16 ***
parking_optionsvalet parking          6.979e+02  4.533e+01  15.397  < 2e-16 ***
cats_allowed                          2.482e+01  5.123e+00   4.845 1.27e-06 ***
smoking_allowed                      -1.248e+02  2.439e+00 -51.183  < 2e-16 ***
comes_furnished                      -4.833e+01  4.971e+00  -9.722  < 2e-16 ***
```

```
Call:
lm(formula = price ~ sqfeet + type + +baths + dogs_allowed +
    electric_vehicle_charge + beds + cats_allowed + smoking_allowed,
    data = data1)

Residuals:
    Min      1Q  Median      3Q     Max
-5942.0  -318.9  -111.2   189.5 12847.3

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)           6.634e+02  4.604e+00 144.107  < 2e-16 ***
sqfeet                5.219e-01  4.408e-03 118.406  < 2e-16 ***
typeassisted living   5.693e+02  4.031e+02   1.412  0.15788
typecondo             3.076e+02  8.635e+00  35.629  < 2e-16 ***
typecottage/cabin     1.297e+02  2.366e+01   5.481 4.24e-08 ***
typeduplex           -6.858e+01  8.856e+00  -7.743 9.71e-15 ***
typeflat              2.138e+02  2.679e+01   7.983 1.43e-15 ***
typehouse            -3.337e+01  4.919e+00  -6.784 1.17e-11 ***
typein-law            2.292e+02  4.759e+01   4.817 1.46e-06 ***
typeland             -6.460e+02  2.550e+02  -2.534  0.01129 *
typeloft              8.257e+01  2.433e+01   3.393  0.00069 ***
typemanufactured     -3.562e+02  9.603e+00 -37.095  < 2e-16 ***
typetownhouse        -7.733e+01  5.377e+00 -14.381  < 2e-16 ***
baths                 1.746e+02  2.854e+00  61.183  < 2e-16 ***
dogs_allowed         -1.366e+01  5.057e+00  -2.702  0.00690 **
electric_vehicle_charge  6.044e+02  9.048e+00  66.799  < 2e-16 ***
beds                 -5.850e+01  2.115e+00 -27.662  < 2e-16 ***
cats_allowed          2.650e+01  5.378e+00   4.929 8.29e-07 ***
smoking_allowed      -1.887e+02  2.521e+00 -74.853  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 570.1 on 233765 degrees of freedom
Multiple R-squared:  0.2142,    Adjusted R-squared:  0.2142
F-statistic:  3541 on 18 and 233765 DF,  p-value: < 2.2e-16
```



Results


```
(Other)                                                                 :233778
                region                                      region_url         price
rochester             :  2864   https://omaha.craigslist.org      :  2023   Min.   :   89
jacksonville          :  2518   https://minneapolis.craigslist.org:  2009   1st Qu.:  824
columbus              :  2112   https://portland.craigslist.org   :  1920   Median : 1064
omaha / council bluffs:  2023   https://sacramento.craigslist.org :  1908   Mean   : 1228
minneapolis / st paul :  2009   https://seattle.craigslist.org    :  1887   3rd Qu.: 1468
jackson               :  1936   https://stlouis.craigslist.org    :  1861   Max.   :14000
(Other)               :220322   (Other)                           :222176
         type            sqfeet          beds           baths         cats_allowed
apartment   :184469   Min.   :   77   Min.   :0.000   Min.   :0.000   Min.   :0.0000
house       : 21985   1st Qu.:  750   1st Qu.:1.000   1st Qu.:1.000   1st Qu.:1.0000
townhouse   : 12783   Median :  950   Median :2.000   Median :1.000   Median :1.0000
condo       :  4643   Mean   : 1008   Mean   :1.913   Mean   :1.479   Mean   :0.7814
duplex      :  4447   3rd Qu.: 1156   3rd Qu.:2.000   3rd Qu.:2.000   3rd Qu.:1.0000
manufactured:  3712   Max.   :11076   Max.   :8.000   Max.   :7.500   Max.   :1.0000
(Other)     :  1745
 dogs_allowed    smoking_allowed  wheelchair_access electric_vehicle_charge
Min.   :0.0000   Min.   :0.0000   Min.   :0.0000    Min.   :0.00000
1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:0.0000    1st Qu.:0.00000
Median :1.0000   Median :1.0000   Median :0.0000    Median :0.00000
Mean   :0.7549   Mean   :0.6435   Mean   :0.1042    Mean   :0.01763
3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000    3rd Qu.:0.00000
Max.   :1.0000   Max.   :1.0000   Max.   :1.0000    Max.   :1.00000

comes_furnished            laundry_options            parking_options       lat
Min.   :0.00000                  :     0   off-street parking:123899   Min.   :-43.53
1st Qu.:0.00000   laundry in bldg  : 31002   attached garage  : 38411   1st Qu.: 33.96
Median :0.00000   laundry on site  : 43754   carport          : 38005   Median : 38.62
Mean   :0.05756   no laundry on site:  3304   detached garage  : 16238   Mean   : 37.90
3rd Qu.:0.00000   w/d hookups      : 54025   street parking   : 15256   3rd Qu.: 41.74
Max.   :1.00000   w/d in unit      :101699   no parking       :  1833   Max.   : 64.99
                                             (Other)          :   142

     long             state
Min.   :-163.89   ca     : 23876
1st Qu.:-105.07   tx     : 15309
Median : -89.39   fl     : 15151
Mean   : -94.21   mi     :  9769
3rd Qu.: -81.57   oh     :  9195
Max.   : 172.63   nc     :  8705
                  (Other):151779
```

```
Subset selection object
Call: regsubsets.formula(price ~ type + sqfeet + beds + baths + cats_allowed +
    dogs_allowed + smoking_allowed + wheelchair_access + electric_vehicle_charge +
    comes_furnished + lat + long + sqfeet * baths + sqfeet *
    beds, data = data1[train, ], nvmax = 50)
24 Variables  (and intercept)
                      Forced in Forced out
typeassisted living         FALSE      FALSE
typecondo                   FALSE      FALSE
typecottage/cabin           FALSE      FALSE
typeduplex                  FALSE      FALSE
typeflat                    FALSE      FALSE
typehouse                   FALSE      FALSE
typein-law                  FALSE      FALSE
typeland                    FALSE      FALSE
typeloft                    FALSE      FALSE
typemanufactured            FALSE      FALSE
typetownhouse               FALSE      FALSE
sqfeet                      FALSE      FALSE
beds                        FALSE      FALSE
baths                       FALSE      FALSE
cats_allowed                FALSE      FALSE
dogs_allowed                FALSE      FALSE
smoking_allowed             FALSE      FALSE
wheelchair_access           FALSE      FALSE
electric_vehicle_charge     FALSE      FALSE
comes_furnished             FALSE      FALSE
lat                         FALSE      FALSE
long                        FALSE      FALSE
sqfeet:baths                FALSE      FALSE
sqfeet:beds                 FALSE      FALSE
1 subsets of each size up to 24
Selection Algorithm: exhaustive
         typeassisted living typecondo typecottage/cabin typeduplex typeflat typehouse
1  ( 1 )  " "                 " "       " "               " "        " "      " "
2  ( 1 )  " "                 " "       " "               " "        " "      " "
3  ( 1 )  " "                 " "       " "               " "        " "      " "
4  ( 1 )  " "                 " "       " "               " "        " "      " "
5  ( 1 )  " "                 " "       " "               " "        " "      " "
6  ( 1 )  " "                 "*"       " "               " "        " "      " "
7  ( 1 )  " "                 "*"       " "               " "        " "      " "
```

```
Statistics by Class:

                     Class: apartment Class: condo Class: duplex Class: house
Sensitivity                    0.9725     0.081406      0.0343277       0.55806
Specificity                    0.4433     0.998667      0.9998104       0.97007
Pos Pred Value                 0.8714     0.554869      0.7795699       0.66116
Neg Pred Value                 0.8059     0.981568      0.9814799       0.95449
Prevalence                     0.7950     0.020006      0.0191623       0.09475
Detection Rate                 0.7731     0.001629      0.0006578       0.05287
Detection Prevalence           0.8872     0.002935      0.0008438       0.07997
Balanced Accuracy              0.7079     0.540036      0.5170690       0.76406
                     Class: manufactured Class: townhouse
Sensitivity                     0.107487          0.25776
Specificity                     0.997451          0.98881
Pos Pred Value                  0.406652          0.57316
Neg Pred Value                  0.985663          0.95807
Prevalence                      0.015996          0.05509
Detection Rate                  0.001719          0.01420
Detection Prevalence            0.004228          0.02477
Balanced Accuracy               0.552469          0.62329
```

Results

- Area
- Housing type
- Number of bathrooms
- Number of beds
- Dogs allowed
- Cats allowed
- Electric_ vehicle charge
- Laundry options
- Parking options
- Area*Number of Beds
- Area*Number of baths

Mean square test error: 4.178

```
Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              1.102e+03  8.037e+00 137.079  < 2e-16 ***
sqfeet                                   2.254e-01  7.442e-03  30.284  < 2e-16 ***
typeassisted living                      4.034e+02  3.810e+02   1.059   0.2896
typecondo                                2.243e+02  8.227e+00  27.257  < 2e-16 ***
typecottage/cabin                        1.707e+02  2.246e+01   7.597 3.04e-14 ***
typeduplex                              -8.325e+01  8.440e+00  -9.863  < 2e-16 ***
typeflat                                 1.830e+02  2.533e+01   7.226 5.00e-13 ***
typehouse                               -1.139e+02  4.932e+00 -23.097  < 2e-16 ***
typein-law                               2.537e+02  4.504e+01   5.633 1.78e-08 ***
typeland                                -5.782e+02  2.409e+02  -2.400   0.0164 *
typeloft                                 9.021e+01  2.301e+01   3.920 8.85e-05 ***
typemanufactured                        -2.056e+02  9.187e+00 -22.379  < 2e-16 ***
typetownhouse                           -7.201e+01  5.149e+00 -13.984  < 2e-16 ***
baths                                    3.764e+01  5.141e+00   7.322 2.46e-13 ***
dogs_allowed                            -4.588e+01  4.832e+00  -9.495  < 2e-16 ***
electric_vehicle_charge                  4.553e+02  8.633e+00  52.743  < 2e-16 ***
beds                                    -4.255e+01  3.486e+00 -12.204  < 2e-16 ***
laundry_optionslaundry on site           1.219e+01  4.048e+00   3.012   0.0026 **
laundry_optionsno laundry on site       -5.581e+01  9.990e+00  -5.586 2.33e-08 ***
laundry_optionsw/d hookups              -9.386e+01  4.123e+00 -22.765  < 2e-16 ***
laundry_optionsw/d in unit               2.254e+02  3.736e+00  60.347  < 2e-16 ***
parking_optionscarport                  -1.605e+02  4.200e+00 -38.218  < 2e-16 ***
parking_optionsdetached garage          -1.440e+02  5.143e+00 -28.004  < 2e-16 ***
parking_optionsno parking               -6.772e+01  1.300e+01  -5.209 1.90e-07 ***
parking_optionsoff-street parking       -3.197e+02  3.576e+00 -89.392  < 2e-16 ***
parking_optionsstreet parking           -2.216e+02  5.395e+00 -41.081  < 2e-16 ***
parking_optionsvalet parking             6.979e+02  4.533e+01  15.397  < 2e-16 ***
cats_allowed                             2.482e+01  5.123e+00   4.845 1.27e-06 ***
smoking_allowed                         -1.248e+02  2.439e+00 -51.183  < 2e-16 ***
comes_furnished                         -4.833e+01  4.971e+00  -9.722  < 2e-16 ***
sqfeet:beds                              4.512e-02  2.588e-03  17.436  < 2e-16 ***
sqfeet:baths                             4.390e-02  3.563e-03  12.322  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 538.7 on 233752 degrees of freedom
Multiple R-squared:  0.2983,    Adjusted R-squared:  0.2983
F-statistic:  3206 on 31 and 233752 DF,  p-value: < 2.2e-16
```

Results


```
k-Nearest Neighbors

11606 samples
   11 predictor
    6 classes: 'apartment', 'condo', 'duplex', 'house', 'manufactured', 'townhouse'

Pre-processing: centered (19), scaled (19)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 10446, 10445, 10445, 10444, 10446, 10444, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   1  0.8269867  0.4998433
   2  0.8157002  0.4621589
   3  0.8325869  0.4845014
   4  0.8350860  0.4830126
   5  0.8377567  0.4797779
   6  0.8378432  0.4742463
   7  0.8375829  0.4680537
   8  0.8368098  0.4637537
   9  0.8368962  0.4553366
  10  0.8362914  0.4507974

Accuracy was used to select the optimal model using the
The final value used for the model was k = 6.
```

```
Statistics by Class:

                     Class: apartment Class: condo Class: duplex Class: house
Sensitivity                    0.9516     0.122676      0.063920      0.54695
Specificity                    0.5321     0.993778      0.996439      0.96776
Pos Pred Value                 0.8875     0.287003      0.259615      0.63969
Neg Pred Value                 0.7393     0.982297      0.981978      0.95329
Prevalence                     0.7950     0.020006      0.019162      0.09475
Detection Rate                 0.7565     0.002454      0.001225      0.05182
Detection Prevalence           0.8524     0.008551      0.004718      0.08101
Balanced Accuracy              0.7419     0.558227      0.530180      0.75735
                     Class: manufactured Class: townhouse
Sensitivity                     0.387975          0.33493
Specificity                     0.993504          0.97648
Pos Pred Value                  0.492618          0.45355
Neg Pred Value                  0.990085          0.96181
Prevalence                      0.015996          0.05509
Detection Rate                  0.006206          0.01845
Detection Prevalence            0.012598          0.04068
Balanced Accuracy               0.690740          0.65570
```

USA housing Listing

Introduction

Data description and preprocessing

Methods

Results

Conclusion and remark

Regression Model

For property owner:

- Loft
- Laundary in unit
- Cat vs Dog
- Parking

For potential tenant:

- Duplex
- Laundary on site

https://reason.com/2019/12/31/virginia-bill-would-end-single-family-only-zoning-in-the-old-dominion/
http://www.oldcentrevillegardens.com/laundry.html

KNN

For potential tenant

https://www.searchenginewatch.com/2018/08/27/the-evolution-of-search-succeeding-in-todays-digital-ecosystem-part-2/

Q&A