STAT8406Project

About the data

My main dataset is from UCI on rental listings from various websites retrieved in December 2019. The downloaded zip file came with two CSV files– one with 10k listings, and one with 100k listings. For the project, I focused on just the smaller dataset. I also merged in 2 other datasets– population from Census Bureau, income from IRS

The main question I wanted to explore is what factors impacted overall rent price. This is also why I merged in a couple other datasets, as I thought it would allow me to get a more complete picture.

Data Wrangling

Note: for purposes of knitted file, the first few code chunks are not run (since they take a while). I exported the results to a CSV which is then loaded further down instead.

Load data

I loaded in the data from the CSV file, and then transformed the time column so I could see when the listings were retrieved. This helped me determine what year I should look for when merging in other datasets.

```
rent_data_10k = read_delim("data/apartments_for_rent_classified_10K.csv", delim=";")
rent_data_10k = rent_data_10k %>% mutate(time = as_datetime(time))
head(rent_data_10k)
```

```
# A tibble: 6 x 22
          id category    title body  amenities bathrooms bedrooms currency fee
       <dbl> <chr>       <chr> <chr> <chr>     <chr>     <chr>    <chr>    <chr>
1 5668626895 housing/re~ Stud~ This~ null      null      0        USD      No
2 5664597177 housing/re~ Stud~ This~ null      null      1        USD      No
3 5668626833 housing/re~ Stud~ This~ null      1         0        USD      No
4 5659918074 housing/re~ Stud~ This~ null      1         0        USD      No
5 5668626759 housing/re~ Stud~ This~ null      null      0        USD      No
6 5667891676 housing/re~ 0 BR~ **RA~ Dishwash~ 1         0        USD      No
# i 13 more variables: has_photo <chr>, pets_allowed <chr>, price <dbl>,
#   price_display <chr>, price_type <chr>, square_feet <dbl>, address <chr>,
#   cityname <chr>, state <chr>, latitude <chr>, longitude <chr>, source <chr>,
#   time <dttm>
```

Merge in population and average income

I used the Google Maps API to get zip codes for each listing based on latitude/longitude, since other datsets I was interested in merging in had zip codes. The R packages for zip codes were extremely slow, so I went with the API route.

```
readRenviron(".env")
get_zip = function(lat,lon) {
  print(paste0(lat,",",lon))
  api_key = Sys.getenv("API_KEY")
  base_url = "https://maps.googleapis.com/maps/api/geocode/json?"
  url = paste0(base_url,"latlng=",lat,",",lon,"&sensor=true&key=",api_key)
  resp = GET(url)
  resp_text = content(resp,as="text")
  json_df = fromJSON(resp_text)
  results = json_df$results
  if (length(results) == 0) {
    return(NULL)
  }
  address_pieces = results$address_components
  max_idx = which.max(sapply(address_pieces, nrow))
  address = address_pieces[[max_idx]]
  postal_code = address %>% filter(types == "postal_code")
  return(postal_code$long_name[1])
}
vectorized_get_zip = Vectorize(get_zip)
rent_data_with_zip = rent_data_10k %>% mutate(
```

```
  Zip_Code = vectorized_get_zip(latitude, longitude)
)
rent_data_with_zip = rent_data_with_zip %>% mutate(Zip_Code = as.character(Zip_Code))
```

Google Maps was able to get zip codes for the majority of listings, but some of them were still null. Luckily, it was only 18 it couldn't find, so I was able to manually set the zip codes based on what was given in the listing description.

```
rent_data_no_zip = rent_data_with_zip %>% filter(is.na(Zip_Code) | Zip_Code == "NULL")
rent_data_zip = rent_data_with_zip %>% filter(Zip_Code != "NULL" & !is.na(Zip_Code))
# manually set zip since there's only 18 left
zips = c(
  "32003",
  "08215",
  "22041",
  "22844",
  "93035",
  "97215",
  "48035",
  "06118",
  "06118",
  "06118",
  "06118",
  "06118",
  "48035",
  "06118",
  "06118",
  "06118",
  "95391",
  "48315"
)
rent_data_no_zip$Zip_Code = zips
rent_data_with_all_zips = rbind(rent_data_zip, rent_data_no_zip)
```

Merge in income data based on zip code

```
# https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2019-zip-c
rent_data_with_income = rent_data_with_all_zips %>% left_join(
  read_csv("data/19zpallnoagi.csv") %>%
    select("ZIPCODE", "A00100"),
  by=c('Zip_Code'='ZIPCODE')
```

```
) %>%
  mutate(agi=A00100) %>%
  select(-c(A00100))
head(rent_data_with_income)
```

Merge in population data based on zip code

```
# https://data.census.gov/table/DECENNIALDHC2020.P1?q=All%205-digit%20ZIP%20Code%20Tabulat
pop_data = read_csv("data/DECENNIALDHC2020.P1-Data.csv") %>%
  select(NAME, P1_001N) %>%
  slice(-1) %>%
  mutate(
    zip = str_sub(NAME, start= -5),
    population = P1_001N
  ) %>%
  select(zip,population)
head(pop_data)
```

Made the column names nicer and saved the new dataset to a csv, so that I wouldn't need to reload zip code data later

```
final_rent_data = rent_data_with_income %>% left_join(pop_data, by=c("Zip_Code"="zip"))
colnames(final_rent_data) = make.names(colnames(final_rent_data))
write.csv(final_rent_data, file="final_rent_data.csv", row.names = FALSE)
```

Now we can work with it (and not rerun all the code from before)

```
loaded_rent_data = read_csv("final_rent_data.csv")
```

```
Rows: 10000 Columns: 25
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (19): category, title, body, amenities, bathrooms, bedrooms, currency, f...
dbl  (6): id, price, square_feet, time, agi, population

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
head(loaded_rent_data)
```

```
# A tibble: 6 x 25
          id category    title body  amenities bathrooms bedrooms currency fee
       <dbl> <chr>       <chr> <chr> <chr>     <chr>     <chr>    <chr>    <chr>
1 5668626895 housing/re~ Stud~ This~ null      null      0        USD      No
2 5664597177 housing/re~ Stud~ This~ null      null      1        USD      No
3 5668626833 housing/re~ Stud~ This~ null      1         0        USD      No
4 5659918074 housing/re~ Stud~ This~ null      1         0        USD      No
5 5668626759 housing/re~ Stud~ This~ null      null      0        USD      No
6 5667891676 housing/re~ 0 BR~ **RA~ Dishwash~ 1         0        USD      No
# i 16 more variables: has_photo <chr>, pets_allowed <chr>, price <dbl>,
#   price_display <chr>, price_type <chr>, square_feet <dbl>, address <chr>,
#   cityname <chr>, state <chr>, latitude <chr>, longitude <chr>, source <chr>,
#   time <dbl>, Zip_Code <chr>, agi <dbl>, population <dbl>
```

Cleaning

```
rent_data = loaded_rent_data %>%
  mutate(
    category=substr(
      category,
      stri_locate_last_fixed(category,"/")[1]+1,
      length(category)
    ),
    bathrooms=ifelse(bathrooms == "null" | is.na(bathrooms), 0, as.numeric(bathrooms)),
    bedrooms=as.numeric(bedrooms),
    fee=(fee=="YES"),
    pets_allowed=ifelse(pets_allowed=="None","null",pets_allowed),
    region=case_when(
      state %in% c("CT","ME","MA","NH","RI","VT")~"New England",
      state %in% c("NJ","NY","PA")~"Middle Atlantic",
      state %in% c("IL","IN","MI","OH","WI")~"East North Central",
      state %in% c("IA","KS","MN","MO","NE","ND","SD")~"West North Central",
      state %in% c("DE","FL","GA","MD","NC","SC","VA","DC","WV")~"South Atlantic",
      state %in% c("AL","KY","MS","TN")~"East South Central",
      state %in% c("AR","LA","OK","TX")~"West South Central",
      state %in% c("AZ","CO","ID","MT","NV","NM","UT","WY")~"Mountain",
      state %in% c("AK","CA","HI","OR","WA")~"Pacific",
    ),
    price=case_when(
      price_type == "Weekly" ~ price*4.5,
      .default=price
```

```
    ),
    has_photo = has_photo %in% c("Thumbnail", "Yes")
  ) %>%
  filter(!is.na(region)) %>%
  select(
    -c(
      price_display,
      title,
      id,
      body,
      address,
      price_type,
      latitude,
      longitude,
      time,
      Zip_Code,
      source,
      cityname,
      state,
      currency
    )
  )
head(rent_data)
```

```
# A tibble: 6 x 12
  category  amenities      bathrooms bedrooms fee   has_photo pets_allowed price
  <chr>     <chr>              <dbl>    <dbl> <lgl> <lgl>     <chr>         <dbl>
1 apartment null                 0        0 FALSE TRUE      null            790
2 apartment null                 0        1 FALSE TRUE      null            425
3 apartment null                 1        0 FALSE TRUE      null           1390
4 apartment null                 1        0 FALSE TRUE      null            925
5 apartment null                 0        0 FALSE TRUE      null            880
6 apartment Dishwasher,El~       1        0 FALSE TRUE      null           2475
# i 4 more variables: square_feet <dbl>, agi <dbl>, population <dbl>,
#   region <chr>
```

Split amenities and pets in separate indicator columns

```
# ChatGPT
split_column = function(df, col) {
  new_df = df %>%
    mutate(id = row_number()) %>%
    separate_rows(!!sym(col), sep = ",") %>%
    mutate(value = TRUE) %>%
    complete(id, !!sym(col), fill = list(value = FALSE)) %>%
    pivot_wider(
      names_from = !!sym(col),
      values_from = value,
      values_fill = list(value = FALSE)
    ) %>%
    select(-id, -null) %>%
    drop_na()
  return(new_df)
}
rent_data = split_column(rent_data, "amenities")
rent_data = split_column(rent_data, "pets_allowed")
```

Remove columns where there is only 1 unique value

```
rent_data = rent_data %>%
  select(where(~ n_distinct(.) > 1))
head(rent_data)
```

```
# A tibble: 6 x 38
  category  bathrooms bedrooms has_photo price square_feet     agi population
  <chr>         <dbl>    <dbl> <lgl>     <dbl>       <dbl>   <dbl>      <dbl>
1 apartment         0        0 TRUE        790         101 3335833      69422
2 apartment         0        1 TRUE        425         106  671972      24895
3 apartment         1        0 TRUE       1390         107 2977581      39331
4 apartment         1        0 TRUE        925         116 2893278      41646
5 apartment         0        0 TRUE        880         125 1494392      25437
6 apartment         1        0 TRUE       2475         130 6847900      49029
# i 30 more variables: region <chr>, AC <lgl>, Alarm <lgl>, Basketball <lgl>,
#   `Cable or Satellite` <lgl>, Clubhouse <lgl>, Dishwasher <lgl>,
#   Doorman <lgl>, Elevator <lgl>, Fireplace <lgl>, `Garbage Disposal` <lgl>,
```

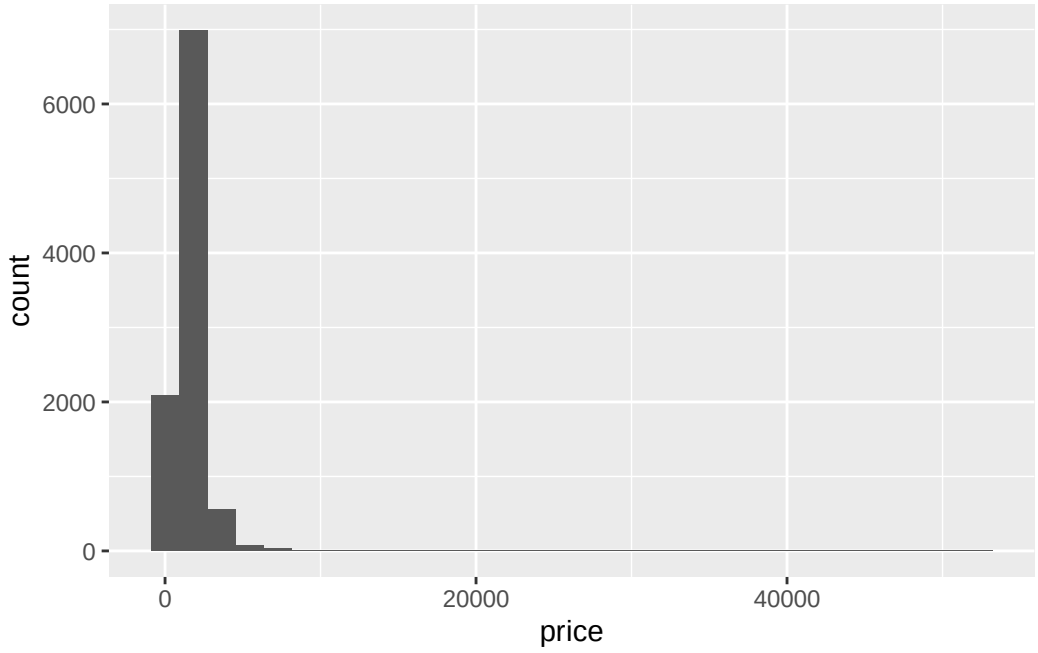
```
#   Gated <lgl>, Golf <lgl>, Gym <lgl>, `Hot Tub` <lgl>,
#   `Internet Access` <lgl>, Luxury <lgl>, Parking <lgl>, `Patio/Deck` <lgl>,
#   Playground <lgl>, Pool <lgl>, Refrigerator <lgl>, Storage <lgl>,
#   Tennis <lgl>, TV <lgl>, View <lgl>, `Washer Dryer` <lgl>, ...
```

Explore Data and Assumptions

Plots for variables of interest

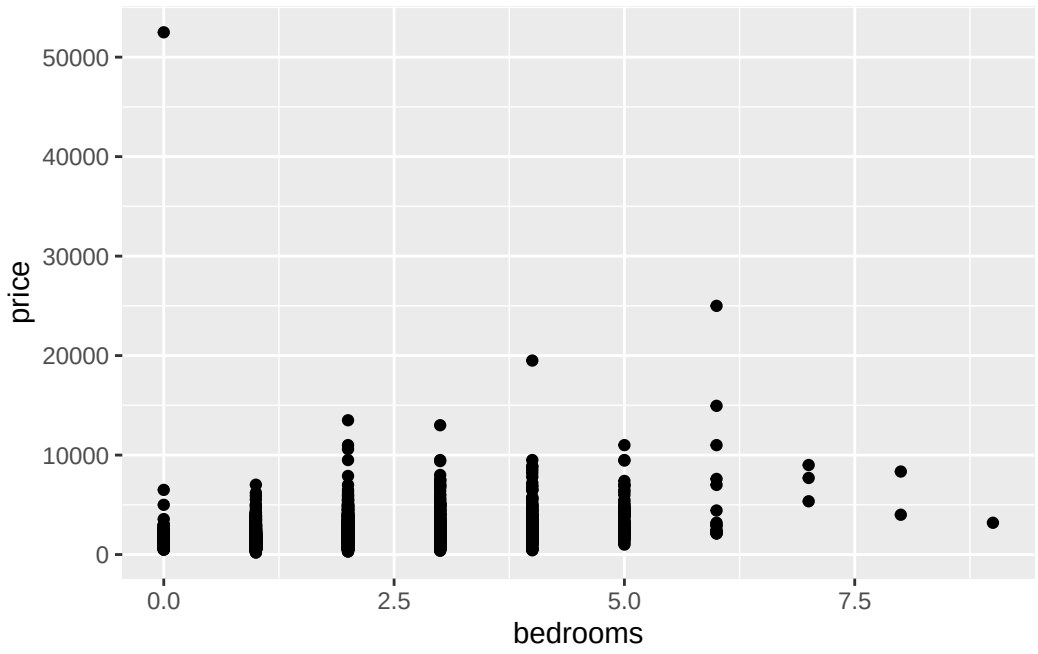
```
ggplot(data=rent_data) + geom_histogram(aes(x=price))
```

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



Data does not appear to be very normal

```
ggplot(data=rent_data) + geom_point(aes(x=bedrooms, y=price))
```



A major outlier, may want to remove

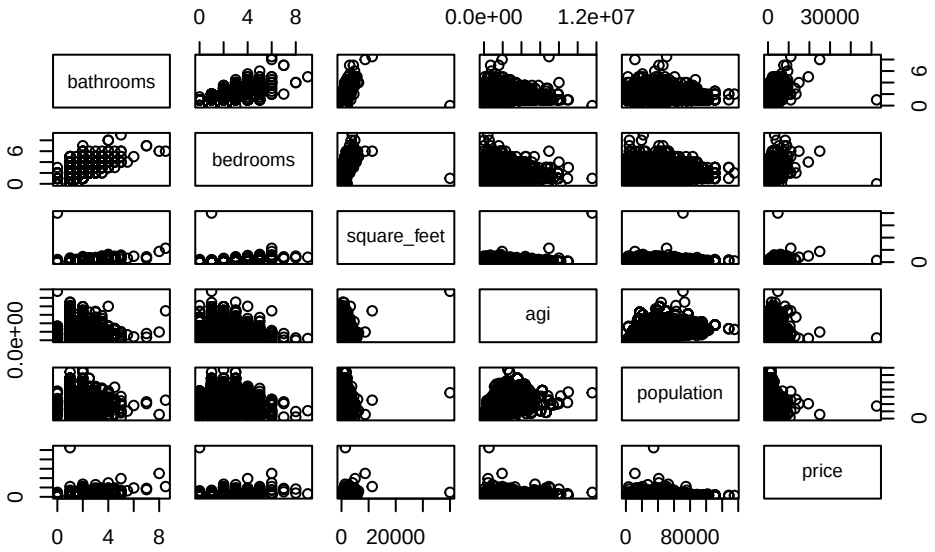
```
rent_data %>% filter(price > 50000)
```

```
# A tibble: 1 x 38
  category  bathrooms bedrooms has_photo price square_feet    agi population
  <chr>         <dbl>    <dbl> <lgl>     <dbl>       <dbl>  <dbl>      <dbl>
1 apartment         1        0 TRUE      52500        1418 538560      34524
# i 30 more variables: region <chr>, AC <lgl>, Alarm <lgl>, Basketball <lgl>,
#   `Cable or Satellite` <lgl>, Clubhouse <lgl>, Dishwasher <lgl>,
#   Doorman <lgl>, Elevator <lgl>, Fireplace <lgl>, `Garbage Disposal` <lgl>,
#   Gated <lgl>, Golf <lgl>, Gym <lgl>, `Hot Tub` <lgl>,
#   `Internet Access` <lgl>, Luxury <lgl>, Parking <lgl>, `Patio/Deck` <lgl>,
#   Playground <lgl>, Pool <lgl>, Refrigerator <lgl>, Storage <lgl>,
#   Tennis <lgl>, TV <lgl>, View <lgl>, `Washer Dryer` <lgl>, ...
```

AGI is very high– either high cost of living or data entry error. It is a large apartment as well however.

Assumptions and Conditions

```
rent_data %>%
  select(
    bathrooms,
    bedrooms,
    square_feet,
    agi,
    population,
    price
  ) %>%
  pairs()
```



Lots of right-skewed relationships. Potentially linear but also appears to be a lot of influential points (especially when focusing on just the relationships with price)

Linear Regression Model

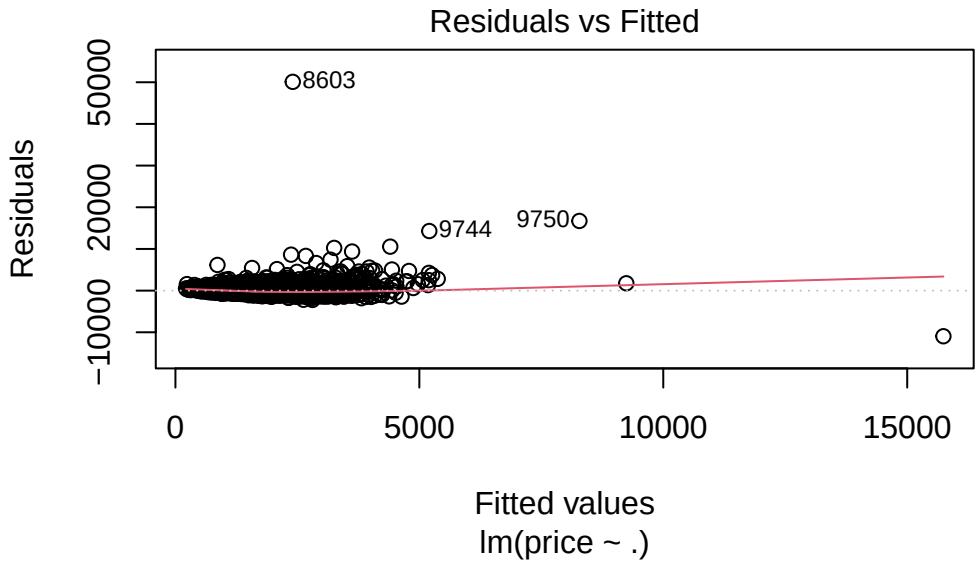
```
rent_lm = lm(price ~ ., data=rent_data)
summary(rent_lm)
```

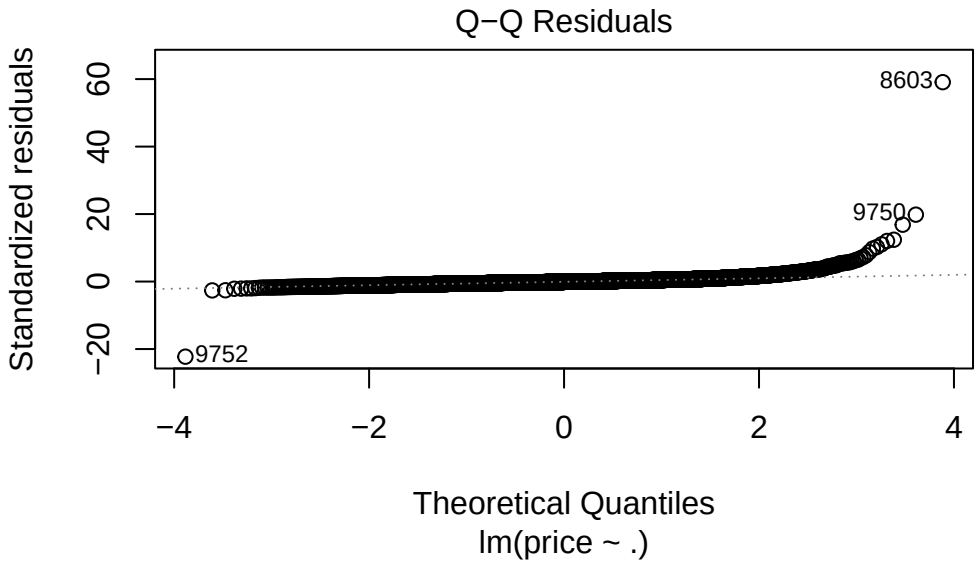
```
Call:
```

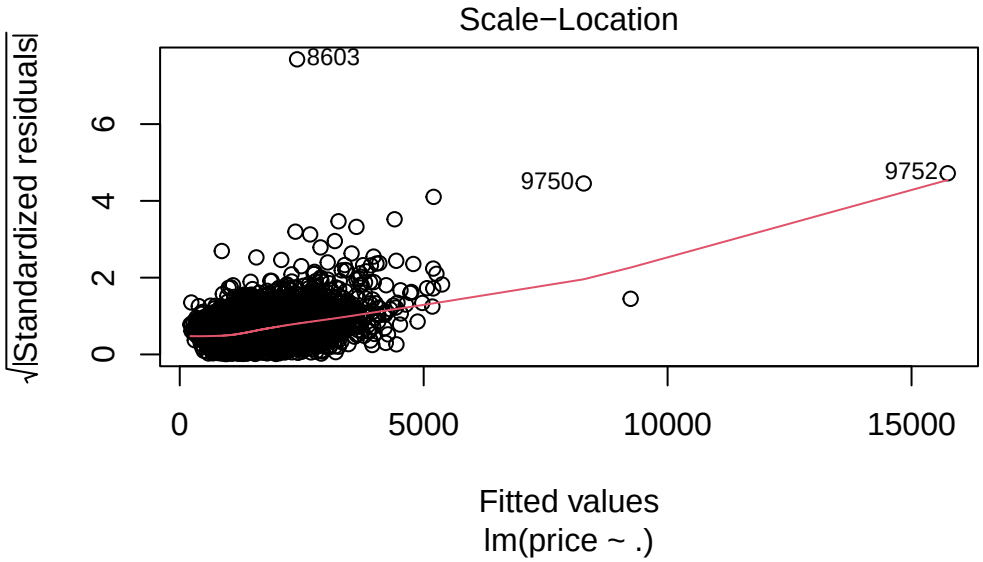
```
lm(formula = price ~ ., data = rent_data)

Residuals:
   Min     1Q Median     3Q    Max
-10952   -351    -71    237  50095

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                2.634e+02  7.507e+01   3.509 0.000452 ***
categoryhome              -5.892e+02  6.063e+02  -0.972 0.331198
categoryshort_term        -3.895e+02  6.040e+02  -0.645 0.519001
bathrooms                  4.673e+02  2.164e+01  21.595  < 2e-16 ***
bedrooms                   1.879e+01  1.390e+01   1.352 0.176404
has_photoTRUE             -2.339e+01  6.449e+01  -0.363 0.716823
square_feet                3.486e-01  1.770e-02  19.693  < 2e-16 ***
agi                        1.709e-04  8.558e-06  19.966  < 2e-16 ***
population                -4.121e-03  5.454e-04  -7.556 4.54e-14 ***
regionEast South Central  -1.641e+02  6.598e+01  -2.487 0.012917 *
regionMiddle Atlantic      4.600e+02  4.289e+01  10.725  < 2e-16 ***
regionMountain             7.209e+00  4.186e+01   0.172 0.863264
regionNew England          6.006e+02  5.174e+01  11.607  < 2e-16 ***
regionPacific              8.664e+02  3.388e+01  25.575  < 2e-16 ***
regionSouth Atlantic      -4.933e+00  3.284e+01  -0.150 0.880587
regionWest North Central  -1.489e+02  3.769e+01  -3.950 7.86e-05 ***
regionWest South Central  -2.298e+02  3.336e+01  -6.888 6.02e-12 ***
ACTRUE                     8.998e+01  4.051e+01   2.221 0.026359 *
AlarmTRUE                  2.596e+01  1.791e+02   0.145 0.884769
BasketballTRUE            -1.779e+02  5.315e+01  -3.348 0.000817 ***
`Cable or Satellite`TRUE   1.816e+01  3.422e+01   0.531 0.595671
ClubhouseTRUE             -1.255e+01  3.132e+01  -0.401 0.688683
DishwasherTRUE             1.182e+01  2.819e+01   0.419 0.675027
DoormanTRUE               -2.096e+02  1.663e+02  -1.261 0.207436
ElevatorTRUE               2.750e+02  3.846e+01   7.151 9.22e-13 ***
FireplaceTRUE             -9.922e+01  3.106e+01  -3.195 0.001403 **
`Garbage Disposal`TRUE    -1.455e+02  3.401e+01  -4.279 1.90e-05 ***
GatedTRUE                 -3.356e+01  4.334e+01  -0.774 0.438742
GolfTRUE                   7.674e+01  1.800e+02   0.426 0.669961
GymTRUE                   -5.790e+00  3.152e+01  -0.184 0.854252
`Hot Tub`TRUE              4.549e+01  5.050e+01   0.901 0.367727
`Internet Access`TRUE      3.385e+01  3.336e+01   1.015 0.310203
LuxuryTRUE                 2.113e+02  2.580e+02   0.819 0.412901
ParkingTRUE                5.708e+00  2.162e+01   0.264 0.791748
`Patio/Deck`TRUE           3.048e+01  2.554e+01   1.194 0.232700
```

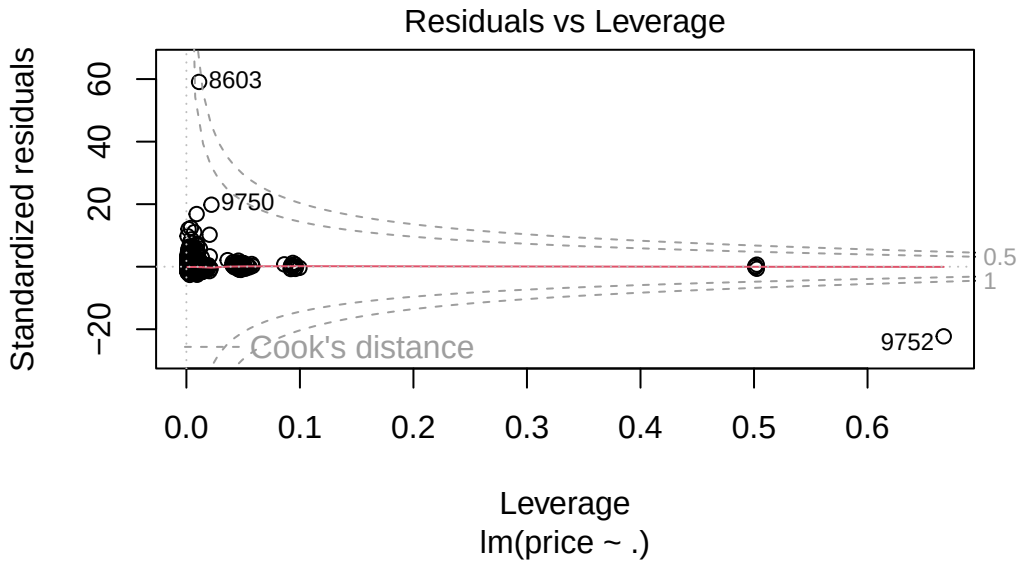
```
PlaygroundTRUE           -2.044e+02  3.617e+01  -5.653 1.62e-08 ***
PoolTRUE                  1.139e+00  2.384e+01   0.048 0.961899
RefrigeratorTRUE          4.211e+01  2.986e+01   1.410 0.158544
StorageTRUE              -3.737e+01  2.680e+01  -1.394 0.163237
TennisTRUE               -8.588e+01  4.484e+01  -1.915 0.055491 .
TVTRUE                    2.147e+02  6.210e+01   3.457 0.000548 ***
ViewTRUE                  1.812e+02  7.270e+01   2.492 0.012703 *
`Washer Dryer`TRUE       -7.571e+01  3.423e+01  -2.212 0.026987 *
`Wood Floors`TRUE         1.762e+02  4.796e+01   3.675 0.000239 ***
CatsTRUE                 -8.789e+00  3.784e+01  -0.232 0.816357
DogsTRUE                 -1.496e+01  3.736e+01  -0.400 0.688923
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 852.5 on 9715 degrees of freedom
Multiple R-squared:  0.385, Adjusted R-squared:  0.3822
F-statistic: 135.2 on 45 and 9715 DF,  p-value: < 2.2e-16
```

```
rent_lm %>% plot()
```




Q−Q Residuals
Standardized residuals
8603
9750
9752
−20
0
20
40
60
−4
−2
0
2
4
Theoretical Quantiles
lm(price ~ .)


Scale−Location
√|Standardized residuals|
8603
9750
9752
0
2
4
6
0
5000
10000
15000
Fitted values
lm(price ~ .)



Based on the first plot (residual vs fitted), there is reason to believe that there are influential points and a possible non-linear relationship. In addition, plot 2 (QQ) shows possible issues with the errors not being constant/normal

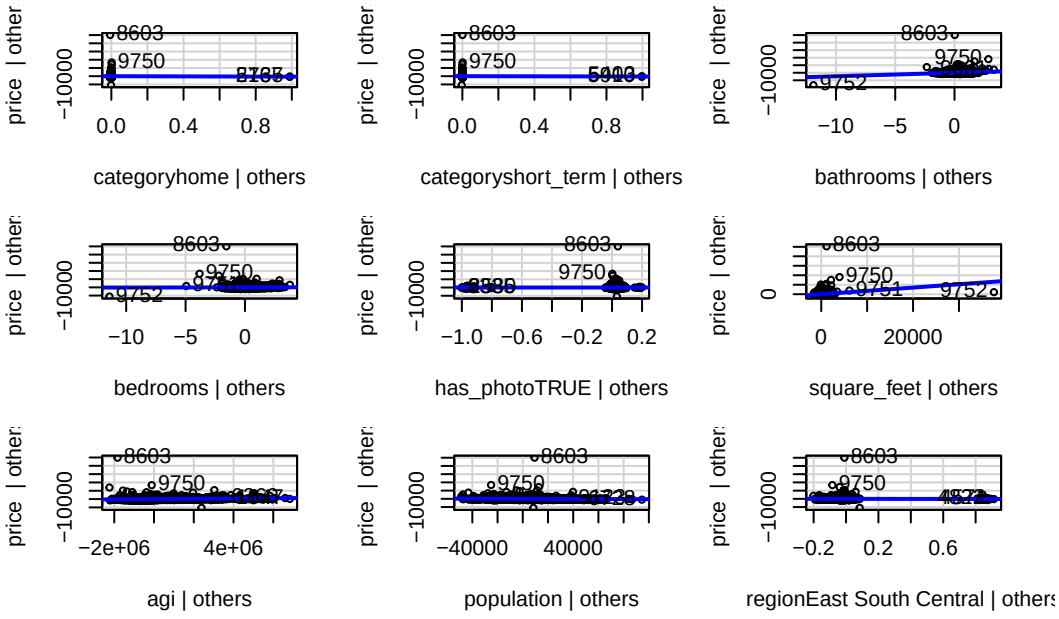
```
vif(rent_lm)
```

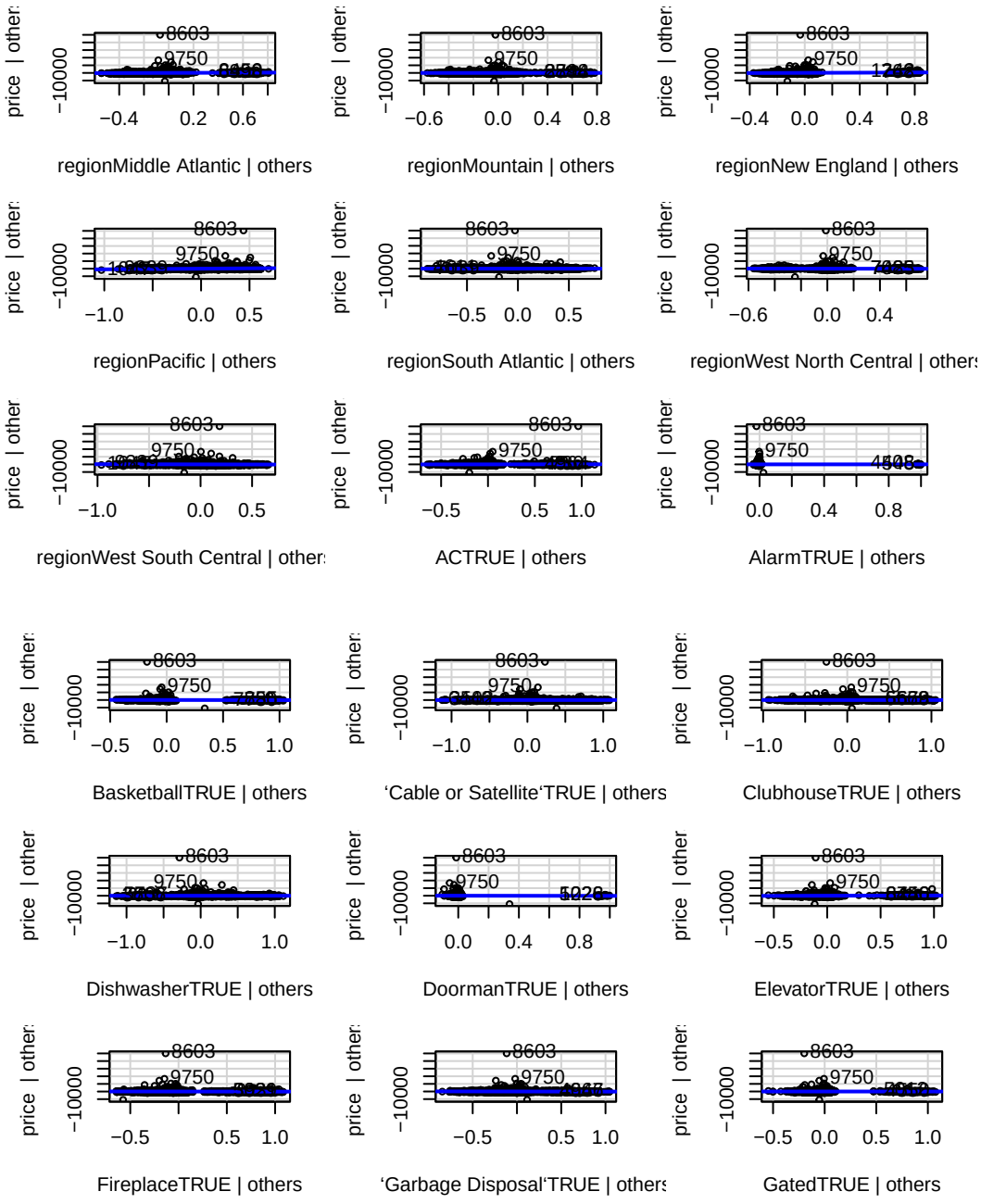
```
                      GVIF Df GVIF^(1/(2*Df))
category          1.014955  2        1.003718
bathrooms         2.426369  1        1.557681
bedrooms          2.308796  1        1.519472
has_photo         1.022061  1        1.010970
square_feet       1.832407  1        1.353664
agi               1.419805  1        1.191556
population        1.234190  1        1.110941
region            1.923986  8        1.041748
AC                1.373805  1        1.172094
Alarm             1.012801  1        1.006380
Basketball        1.184718  1        1.088447
`Cable or Satellite` 2.214574  1        1.488144
Clubhouse         1.526023  1        1.235323
Dishwasher        2.359821  1        1.536171
Doorman           1.061989  1        1.030529
Elevator          1.204576  1        1.097532
Fireplace         1.252813  1        1.119291
`Garbage Disposal`   1.666788  1        1.291041
```

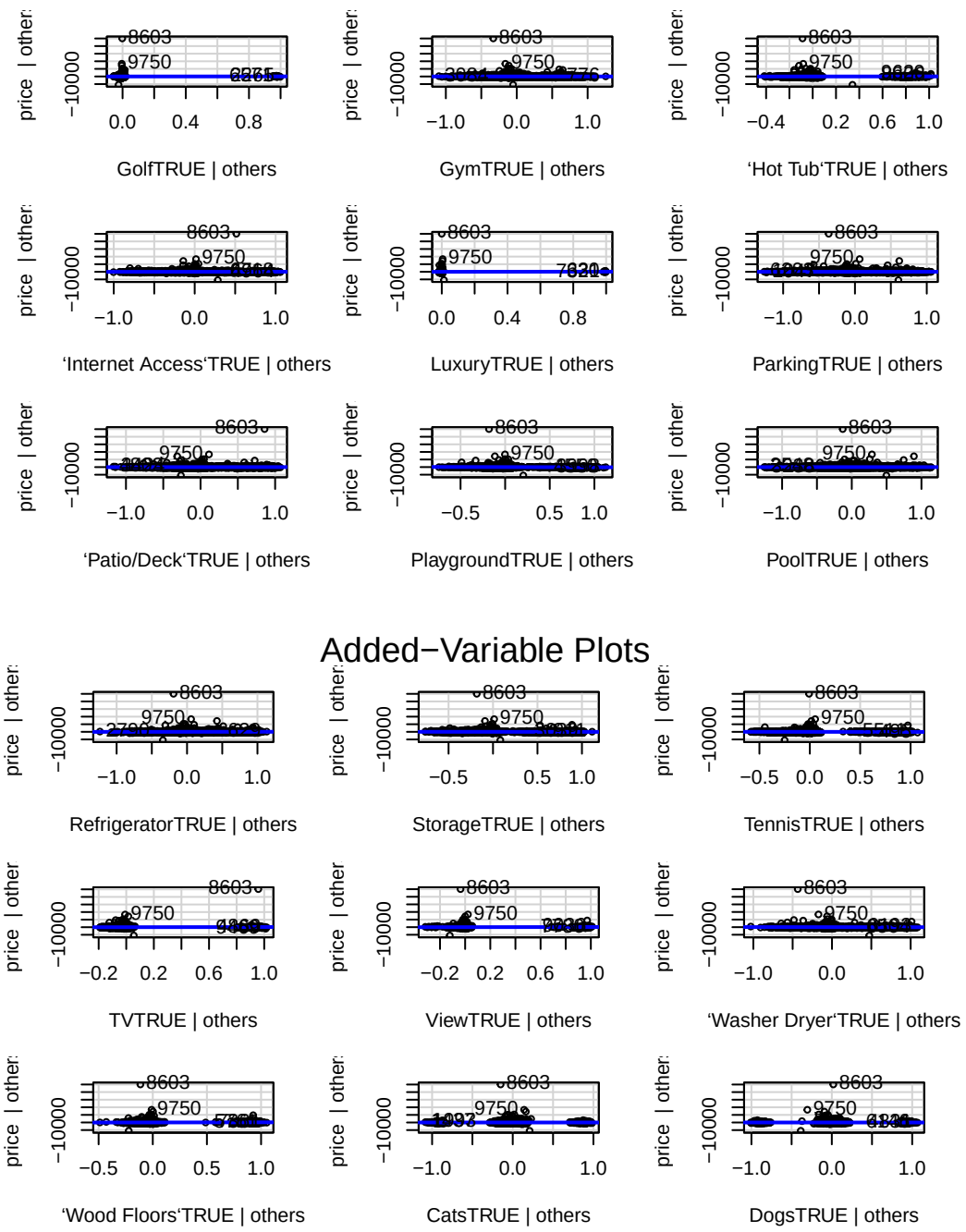
```
Gated                1.146948  1        1.070957
Golf                 1.023424  1        1.011644
Gym                  1.698770  1        1.303369
`Hot Tub`            1.154613  1        1.074529
`Internet Access`    1.862015  1        1.364557
Luxury               1.006286  1        1.003138
Parking              1.475008  1        1.214499
`Patio/Deck`         1.644070  1        1.282213
Playground           1.284009  1        1.133141
Pool                 1.682614  1        1.297156
Refrigerator         2.589022  1        1.609044
Storage              1.266834  1        1.125537
Tennis               1.250171  1        1.118110
TV                   1.065050  1        1.032013
View                 1.059818  1        1.029475
`Washer Dryer`       1.534373  1        1.238698
`Wood Floors`        1.088398  1        1.043263
Cats                 4.706311  1        2.169403
Dogs                 4.658809  1        2.158427
```

All of the VIF's are pretty low. Cats and dogs are above 3, so that's something to take note of, but not high enough to be cause for concern

```
rent_lm %>% avPlots(ask = F)
```

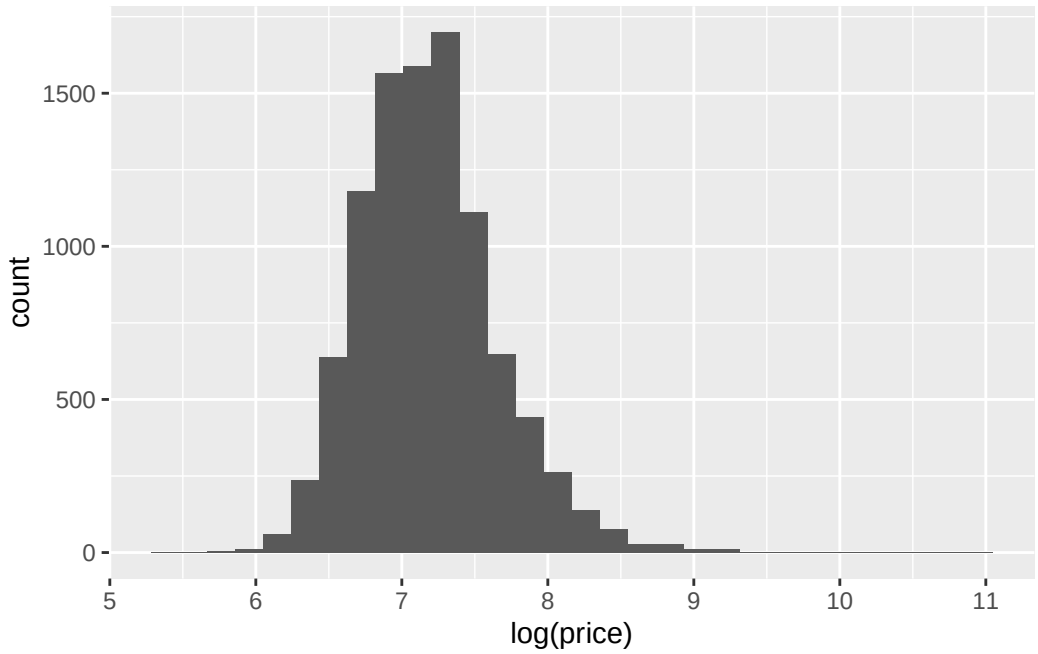



price | others
−10000
8603
9750
regionMiddle Atlantic | others
regionMountain | others
regionNew England | others
regionPacific | others
regionSouth Atlantic | others
regionWest North Central | others
regionWest South Central | others
ACTRUE | others
AlarmTRUE | others
BasketballTRUE | others
‘Cable or Satellite‘TRUE | others
ClubhouseTRUE | others
DishwasherTRUE | others
DoormanTRUE | others
ElevatorTRUE | others
FireplaceTRUE | others
‘Garbage Disposal‘TRUE | others
GatedTRUE | others




```
ggplot(data=rent_data) + geom_histogram(aes(x=log(price)))
```

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



Data looks significantly more normal

Influential points

```
# https://stackoverflow.com/questions/70149293/error-in-view-cannot-coerce-class-infl-to-a
infl = influence.measures(rent_lm)
infl_df = as.data.frame(infl[["infmat"]]) %>%
  dplyr::mutate(inf = ifelse(row_number() %in% unname(which(
    apply(infl$is.inf, 1, any)
  )), "*", ""))
row_nums = which(infl_df$inf != "")
length(row_nums)
```

```
[1] 648
```

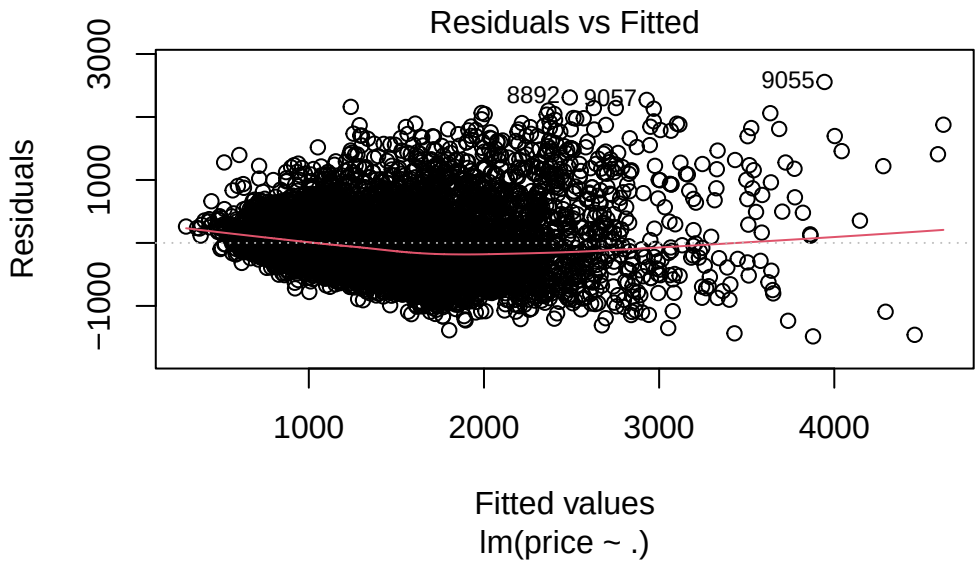
```
rent_data_no_infl = rent_data %>% slice(-row_nums) %>% select(where(~ n_distinct(.) > 1))
rent_lm_no_infl = lm(price ~ ., data=rent_data_no_infl)
summary(rent_lm_no_infl)
```

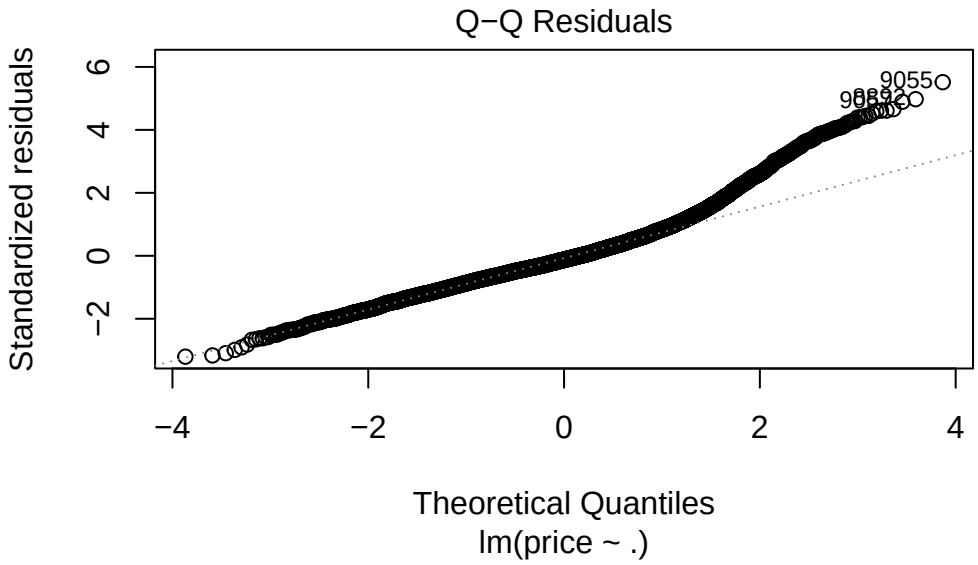
```
Call:
lm(formula = price ~ ., data = rent_data_no_infl)

Residuals:
     Min       1Q   Median       3Q      Max
-1483.56  -289.62   -58.85   222.09  2556.00

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               5.021e+02  4.950e+01  10.144  < 2e-16 ***
bathrooms                 1.810e+02  1.403e+01  12.902  < 2e-16 ***
bedrooms                 -1.288e+00  8.520e+00  -0.151 0.879870
has_photoTRUE            -8.049e+01  4.438e+01  -1.814 0.069752 .
square_feet               5.367e-01  1.786e-02  30.051  < 2e-16 ***
agi                       1.560e-04  4.921e-06  31.703  < 2e-16 ***
population               -3.168e-03  3.095e-04 -10.234  < 2e-16 ***
regionEast South Central -1.124e+02  3.872e+01  -2.902 0.003721 **
regionMiddle Atlantic     4.383e+02  2.433e+01  18.015  < 2e-16 ***
regionMountain            3.176e+01  2.383e+01   1.333 0.182713
regionNew England         5.810e+02  3.009e+01  19.311  < 2e-16 ***
regionPacific             7.249e+02  1.909e+01  37.965  < 2e-16 ***
regionSouth Atlantic      3.502e+01  1.840e+01   1.904 0.056999 .
regionWest North Central -1.357e+02  2.097e+01  -6.470 1.03e-10 ***
regionWest South Central -2.080e+02  1.866e+01 -11.144  < 2e-16 ***
ACTRUE                   -6.671e+01  2.480e+01  -2.690 0.007162 **
BasketballTRUE           -8.747e+01  3.655e+01  -2.393 0.016726 *
`Cable or Satellite`TRUE -1.467e+01  2.023e+01  -0.725 0.468473
ClubhouseTRUE             2.222e+01  1.853e+01   1.199 0.230588
DishwasherTRUE            3.596e+01  1.597e+01   2.252 0.024376 *
ElevatorTRUE              2.735e+02  2.287e+01  11.958  < 2e-16 ***
FireplaceTRUE            -9.948e+01  1.836e+01  -5.417 6.21e-08 ***
`Garbage Disposal`TRUE   -9.186e+01  1.965e+01  -4.674 2.99e-06 ***
GatedTRUE                 4.064e+00  2.712e+01   0.150 0.880870
GymTRUE                   3.287e+01  1.844e+01   1.782 0.074727 .
`Hot Tub`TRUE             9.580e+01  3.316e+01   2.889 0.003877 **
`Internet Access`TRUE     2.196e+01  1.981e+01   1.109 0.267450
ParkingTRUE               3.057e+01  1.233e+01   2.479 0.013208 *
`Patio/Deck`TRUE          2.391e-01  1.468e+01   0.016 0.987005
PlaygroundTRUE           -1.576e+02  2.143e+01  -7.353 2.11e-13 ***
PoolTRUE                 -1.026e+01  1.369e+01  -0.750 0.453312
RefrigeratorTRUE         -1.901e+01  1.712e+01  -1.110 0.266933
StorageTRUE              -2.340e+01  1.559e+01  -1.501 0.133409
```

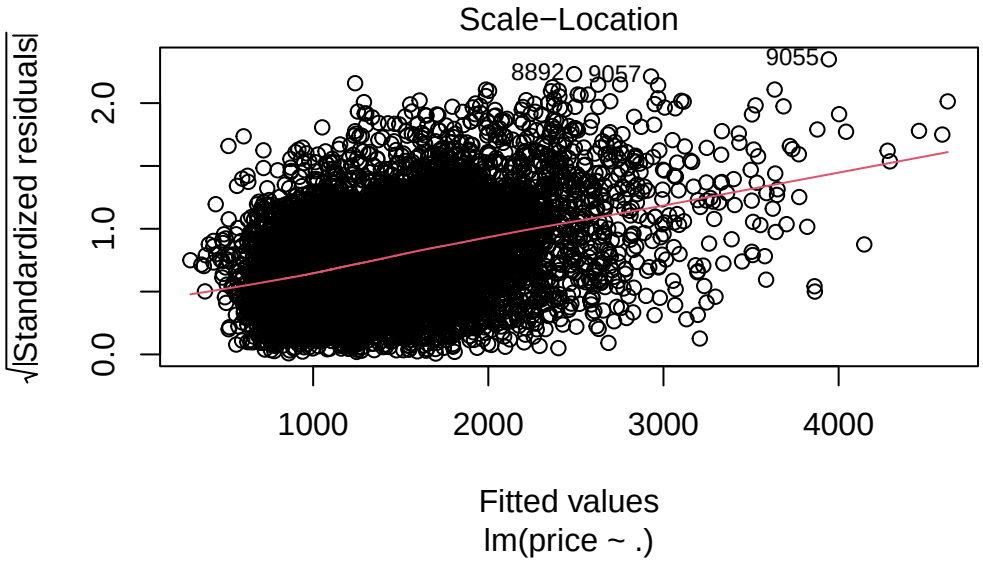
```
TennisTRUE                -9.903e+01  2.773e+01  -3.571 0.000358 ***
TVTRUE                    -9.773e+01  4.954e+01  -1.973 0.048568 *
ViewTRUE                   7.689e+01  7.357e+01   1.045 0.296029
`Washer Dryer`TRUE        -8.715e+00  2.016e+01  -0.432 0.665513
`Wood Floors`TRUE          1.580e+02  2.918e+01   5.415 6.28e-08 ***
CatsTRUE                   1.824e+01  2.119e+01   0.861 0.389420
DogsTRUE                  -9.835e+00  2.097e+01  -0.469 0.639101
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 465.2 on 9073 degrees of freedom
Multiple R-squared:  0.5579,    Adjusted R-squared:  0.556
F-statistic: 293.6 on 39 and 9073 DF,  p-value: < 2.2e-16
```

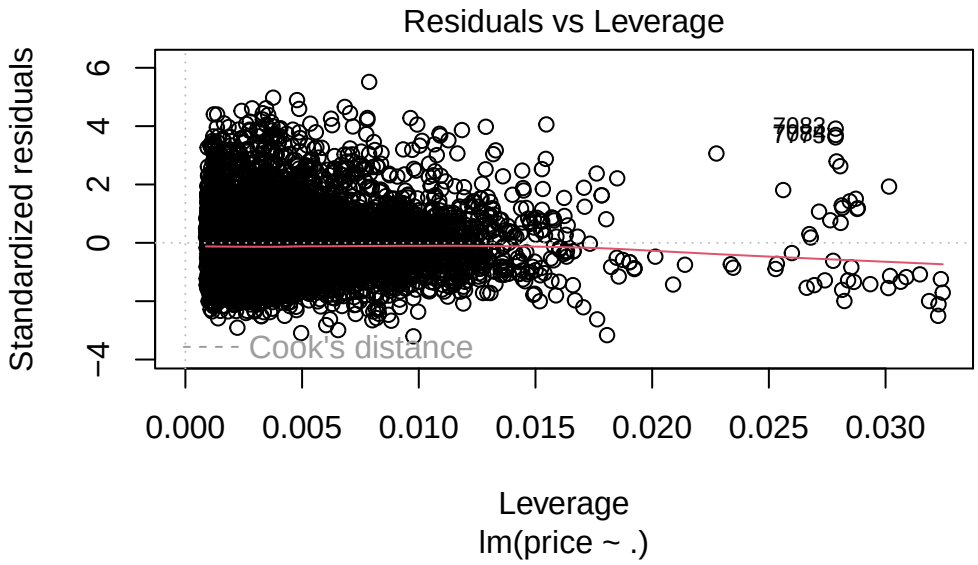
Difference in R-squareds after removing influential points (0.385 vs 0.56 is explained by the model). However there were 648 points removed, and as seen below, still violation of assumptions. We'll transform price and proceed.

```
rent_lm_no_infl %>% plot()
```




Q−Q Residuals
Standardized residuals
9055
−4
−2
0
2
4
6
Theoretical Quantiles
lm(price ~ .)


Scale−Location
√|Standardized residuals|
8892
9057
9055
0.0
1.0
2.0
1000
2000
3000
4000
Fitted values
lm(price ~ .)



Transformation on y

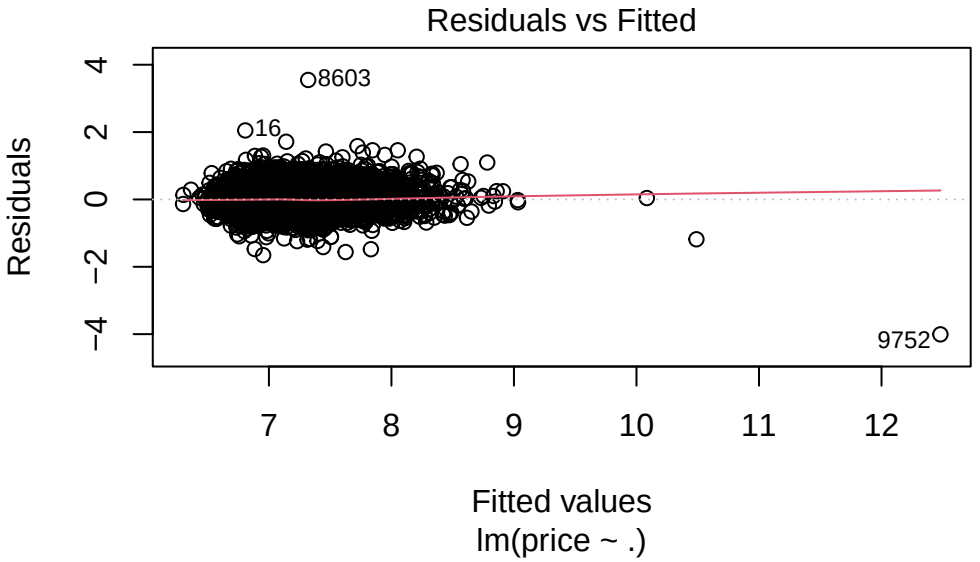
```
rent_data_log = rent_data %>% mutate(price = log(price))
rent_log_lm = lm(price ~ ., data=rent_data_log)
summary(rent_log_lm)
```

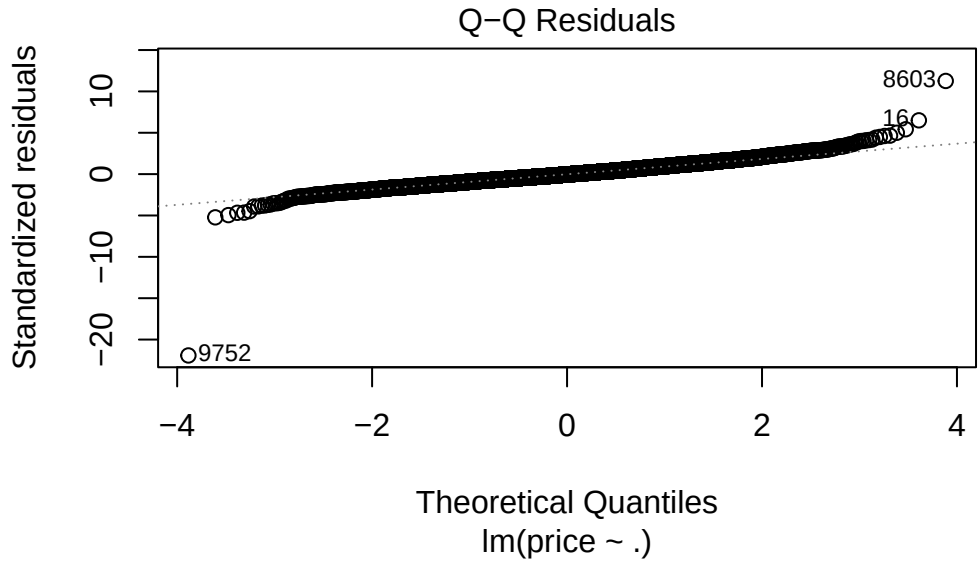
```
Call:
lm(formula = price ~ ., data = rent_data_log)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0064 -0.2022 -0.0139  0.1920  3.5482

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              6.532e+00  2.789e-02 234.206  < 2e-16 ***
categoryhome            -6.692e-01  2.253e-01  -2.971 0.002976 **
categoryshort_term      -5.557e-01  2.244e-01  -2.476 0.013290 *
bathrooms                1.931e-01  8.040e-03  24.015  < 2e-16 ***
bedrooms                 5.714e-02  5.162e-03  11.069  < 2e-16 ***
has_photoTRUE           -3.902e-02  2.396e-02  -1.629 0.103445
square_feet              1.220e-04  6.577e-06  18.556  < 2e-16 ***
agi                      8.984e-08  3.180e-09  28.255  < 2e-16 ***
```

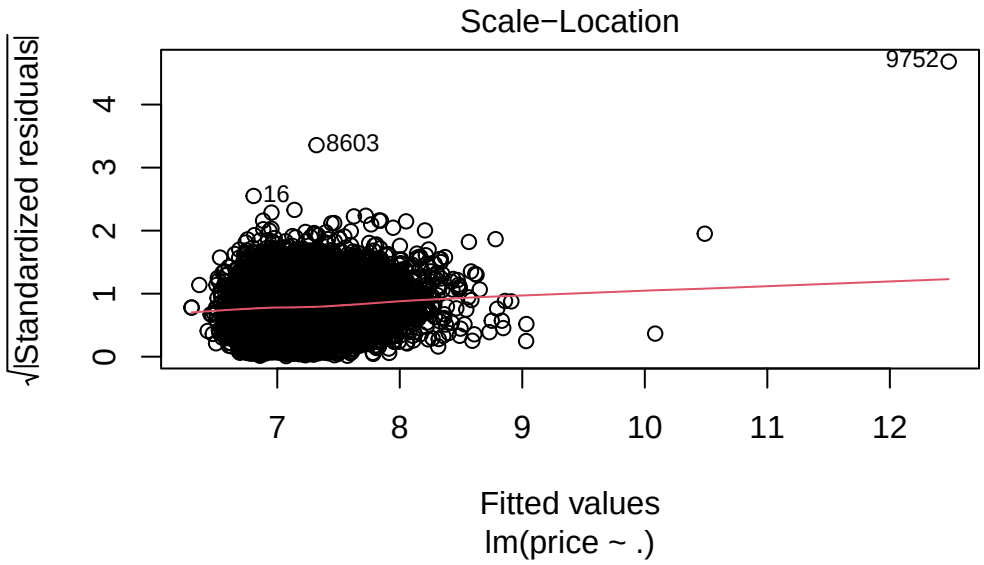
```
population                -1.440e-06  2.026e-07  -7.106 1.28e-12 ***
regionEast South Central  -7.997e-02  2.451e-02  -3.262 0.001109 **
regionMiddle Atlantic      3.387e-01  1.594e-02  21.254  < 2e-16 ***
regionMountain             5.714e-02  1.555e-02   3.674 0.000240 ***
regionNew England          4.007e-01  1.922e-02  20.843  < 2e-16 ***
regionPacific              4.827e-01  1.259e-02  38.349  < 2e-16 ***
regionSouth Atlantic       8.055e-02  1.220e-02   6.602 4.28e-11 ***
regionWest North Central  -9.891e-02  1.400e-02  -7.064 1.72e-12 ***
regionWest South Central  -1.043e-01  1.240e-02  -8.411  < 2e-16 ***
ACTRUE                    -3.473e-02  1.505e-02  -2.307 0.021060 *
AlarmTRUE                  5.267e-02  6.654e-02   0.791 0.428675
BasketballTRUE            -8.616e-02  1.975e-02  -4.364 1.29e-05 ***
`Cable or Satellite`TRUE  -1.892e-02  1.271e-02  -1.489 0.136649
ClubhouseTRUE              7.113e-03  1.164e-02   0.611 0.541049
DishwasherTRUE             3.312e-02  1.047e-02   3.162 0.001570 **
DoormanTRUE               -4.870e-02  6.177e-02  -0.788 0.430462
ElevatorTRUE               2.121e-01  1.429e-02  14.846  < 2e-16 ***
FireplaceTRUE             -3.612e-02  1.154e-02  -3.131 0.001748 **
`Garbage Disposal`TRUE    -7.232e-02  1.264e-02  -5.723 1.08e-08 ***
GatedTRUE                 -2.290e-03  1.610e-02  -0.142 0.886902
GolfTRUE                   7.864e-02  6.689e-02   1.176 0.239795
GymTRUE                    4.670e-02  1.171e-02   3.988 6.70e-05 ***
`Hot Tub`TRUE              6.163e-02  1.876e-02   3.285 0.001023 **
`Internet Access`TRUE      1.754e-02  1.239e-02   1.415 0.157039
LuxuryTRUE                 1.360e-01  9.585e-02   1.418 0.156082
ParkingTRUE                1.767e-02  8.032e-03   2.200 0.027826 *
`Patio/Deck`TRUE           1.866e-03  9.489e-03   0.197 0.844151
PlaygroundTRUE            -1.242e-01  1.344e-02  -9.240  < 2e-16 ***
PoolTRUE                  -1.642e-02  8.858e-03  -1.854 0.063790 .
RefrigeratorTRUE          -3.708e-02  1.109e-02  -3.343 0.000833 ***
StorageTRUE               -8.737e-03  9.957e-03  -0.877 0.380239
TennisTRUE                -5.602e-02  1.666e-02  -3.363 0.000775 ***
TVTRUE                     1.402e-02  2.307e-02   0.608 0.543371
ViewTRUE                   1.134e-01  2.701e-02   4.200 2.69e-05 ***
`Washer Dryer`TRUE        -1.841e-02  1.272e-02  -1.448 0.147667
`Wood Floors`TRUE          1.076e-01  1.782e-02   6.038 1.62e-09 ***
CatsTRUE                   1.145e-02  1.406e-02   0.814 0.415627
DogsTRUE                   1.323e-02  1.388e-02   0.953 0.340573
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3167 on 9715 degrees of freedom
Multiple R-squared:  0.5452,    Adjusted R-squared:  0.5431
```

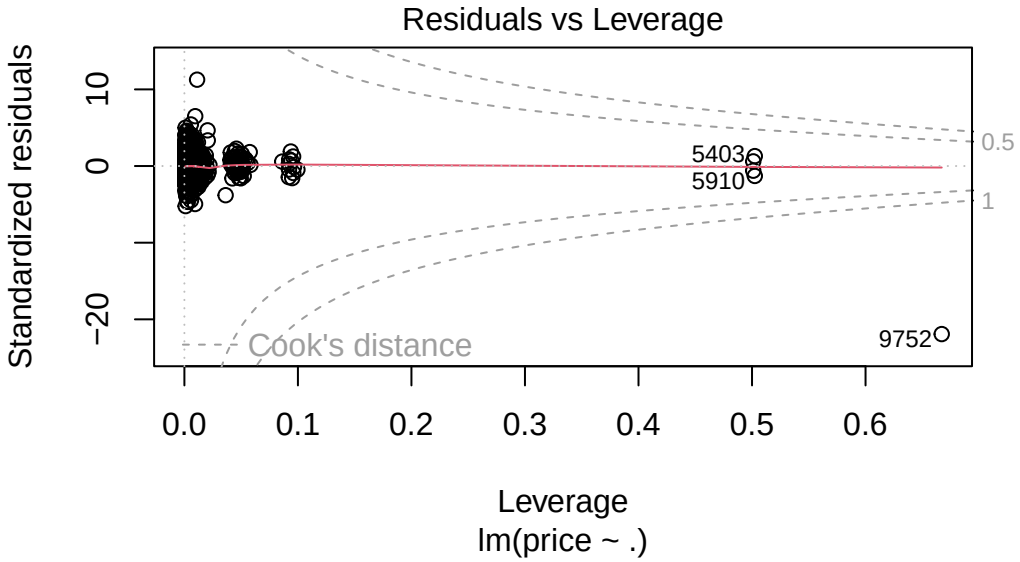
```
F-statistic: 258.8 on 45 and 9715 DF,  p-value: < 2.2e-16
```

```
rent_log_lm %>% plot()
```









Assumptions are definitely closer to being met. A couple possible influential points but significantly more normal/constantly errors.

Just for interest, I decided to look at influential points again. This time, there were 700+. I'll leave them in since our assumptions look closer to being met

```
infl = influence.measures(rent_log_lm)
infl_df = as.data.frame(infl[["infmat"]]) %>%
```

```
  dplyr::mutate(inf = ifelse(row_number() %in% unname(which(
    apply(infl$is.inf, 1, any)
  )), "*", ""))
row_nums = which(infl_df$inf != "")
length(row_nums)
```

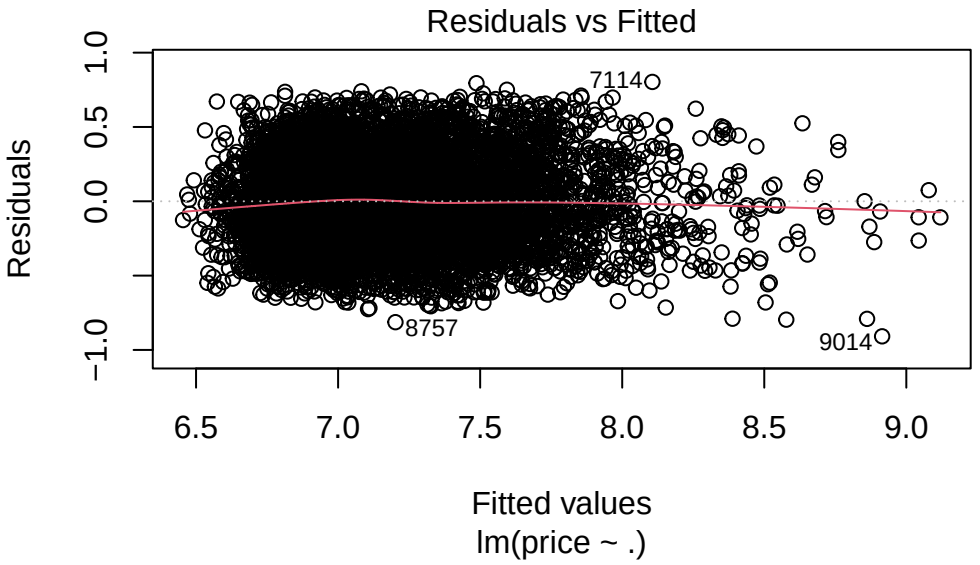
```
[1] 737
```

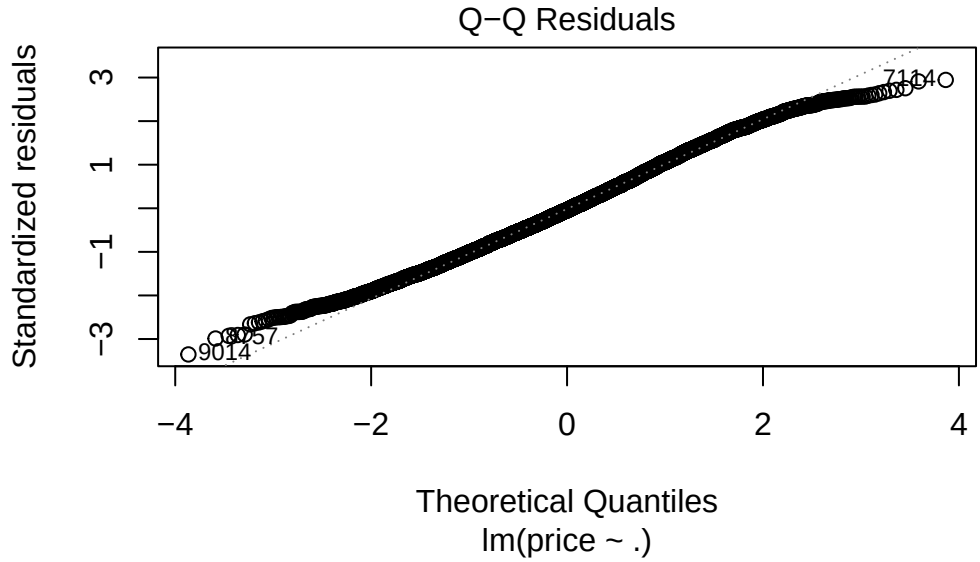
```
rent_data_no_infl = rent_data_log %>% slice(-row_nums) %>% select(where(~ n_distinct(.) >
rent_lm_no_infl = lm(price ~ ., data=rent_data_no_infl)
summary(rent_lm_no_infl)
```

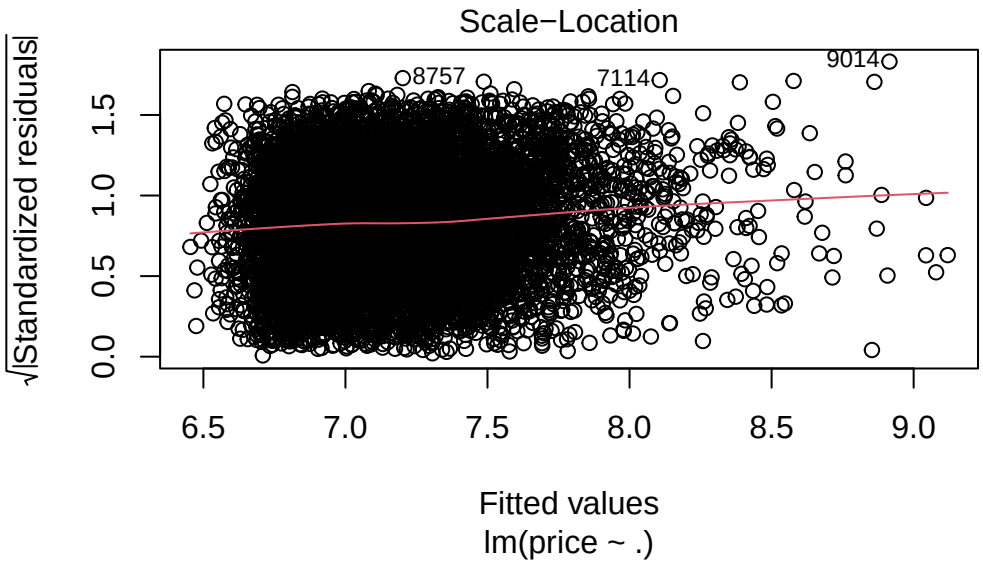
```
Call:
lm(formula = price ~ ., data = rent_data_no_infl)

Residuals:
     Min       1Q   Median       3Q      Max
-0.90850 -0.19288 -0.01127  0.18550  0.80300

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                6.518e+00  2.825e-02 230.697  < 2e-16 ***
bathrooms                  1.129e-01  8.271e-03  13.648  < 2e-16 ***
bedrooms                   2.606e-02  4.975e-03   5.239 1.66e-07 ***
has_photoTRUE             -3.535e-02  2.521e-02  -1.402 0.160874
square_feet                2.812e-04  1.044e-05  26.938  < 2e-16 ***
agi                        9.187e-08  2.884e-09  31.851  < 2e-16 ***
population                -1.550e-06  1.821e-07  -8.514  < 2e-16 ***
regionEast South Central  -6.108e-02  2.289e-02  -2.669 0.007630 **
regionMiddle Atlantic      3.583e-01  1.445e-02  24.801  < 2e-16 ***
regionMountain             7.757e-02  1.394e-02   5.562 2.74e-08 ***
regionNew England          4.097e-01  1.761e-02  23.267  < 2e-16 ***
regionPacific              4.986e-01  1.136e-02  43.897  < 2e-16 ***
regionSouth Atlantic       9.778e-02  1.090e-02   8.970  < 2e-16 ***
regionWest North Central  -9.393e-02  1.244e-02  -7.552 4.70e-14 ***
regionWest South Central  -8.858e-02  1.105e-02  -8.013 1.26e-15 ***
ACTRUE                    -5.143e-02  1.458e-02  -3.528 0.000420 ***
BasketballTRUE            -6.741e-02  2.082e-02  -3.238 0.001208 **
```

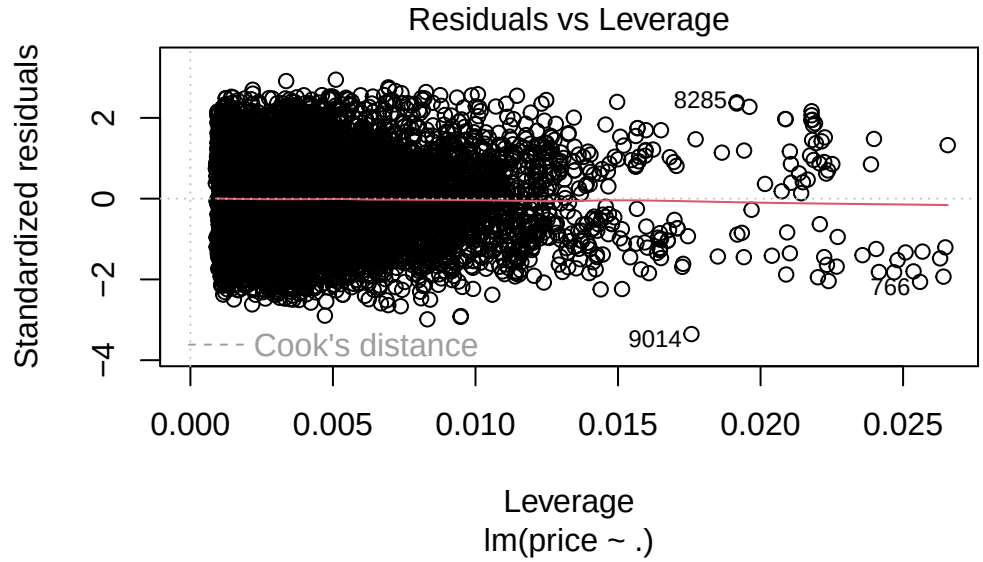
```
`Cable or Satellite`TRUE -8.773e-03  1.180e-02  -0.744 0.457036
ClubhouseTRUE             1.989e-02  1.081e-02   1.840 0.065807 .
DishwasherTRUE            4.498e-02  9.532e-03   4.718 2.41e-06 ***
ElevatorTRUE              2.097e-01  1.346e-02  15.581  < 2e-16 ***
FireplaceTRUE            -5.263e-02  1.075e-02  -4.896 9.95e-07 ***
`Garbage Disposal`TRUE   -6.743e-02  1.155e-02  -5.837 5.49e-09 ***
GatedTRUE                -5.242e-03  1.558e-02  -0.336 0.736519
GymTRUE                   4.860e-02  1.076e-02   4.515 6.40e-06 ***
`Hot Tub`TRUE             8.540e-02  1.901e-02   4.492 7.14e-06 ***
`Internet Access`TRUE     2.431e-02  1.158e-02   2.100 0.035720 *
ParkingTRUE               3.265e-02  7.245e-03   4.506 6.70e-06 ***
`Patio/Deck`TRUE         -1.335e-03  8.607e-03  -0.155 0.876738
PlaygroundTRUE           -1.161e-01  1.244e-02  -9.330  < 2e-16 ***
PoolTRUE                 -1.905e-02  8.046e-03  -2.368 0.017912 *
RefrigeratorTRUE         -5.144e-02  1.020e-02  -5.041 4.72e-07 ***
StorageTRUE              -6.428e-03  9.108e-03  -0.706 0.480396
TennisTRUE               -5.878e-02  1.595e-02  -3.686 0.000229 ***
TVTRUE                   -6.014e-03  2.646e-02  -0.227 0.820177
ViewTRUE                  1.468e-01  3.738e-02   3.928 8.62e-05 ***
`Washer Dryer`TRUE       -7.614e-03  1.175e-02  -0.648 0.516899
`Wood Floors`TRUE         9.881e-02  1.717e-02   5.755 8.96e-09 ***
CatsTRUE                  1.758e-02  1.250e-02   1.406 0.159784
DogsTRUE                  4.250e-03  1.236e-02   0.344 0.731053
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2733 on 8984 degrees of freedom
Multiple R-squared:  0.6084,    Adjusted R-squared:  0.6067
F-statistic: 357.9 on 39 and 8984 DF,  p-value: < 2.2e-16
```

```
rent_lm_no_infl %>% plot()
```

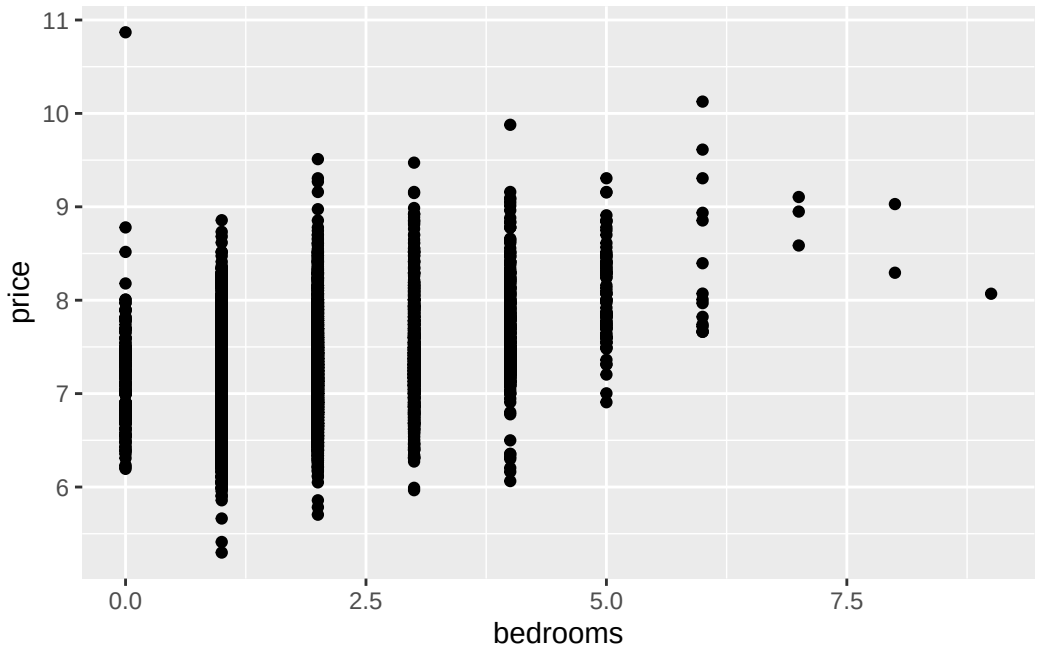

Residuals vs Fitted
Residuals
1.0
0.5
0.0
−1.0
7114
8757
9014
6.5
7.0
7.5
8.0
8.5
9.0
Fitted values
lm(price ~ .)


Q−Q Residuals
Standardized residuals
3
1
−1
−3
7114
9014
−4
−2
0
2
4
Theoretical Quantiles
lm(price ~ .)



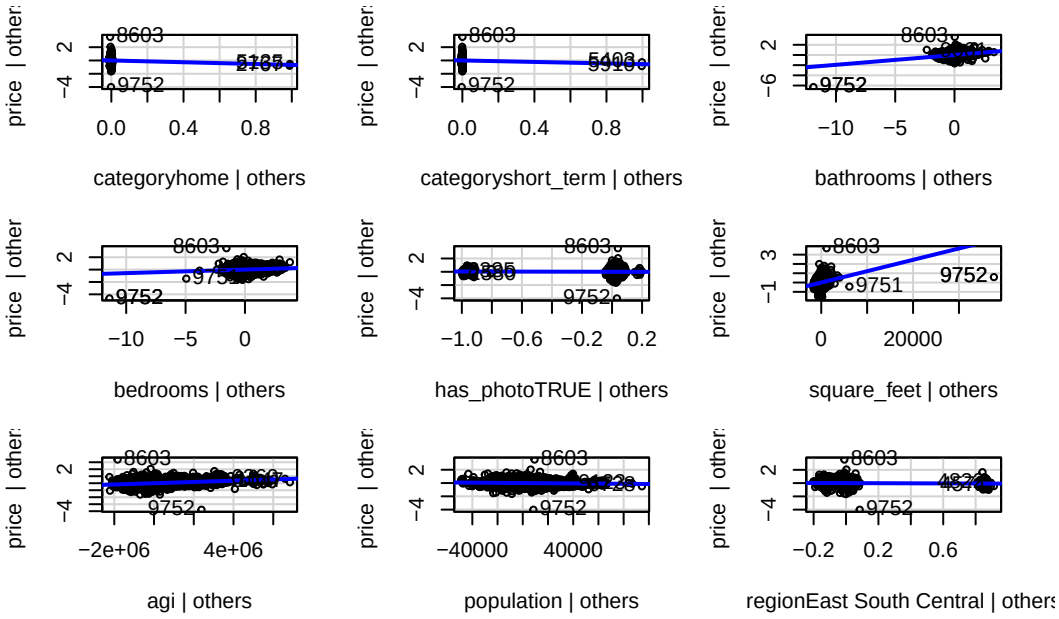


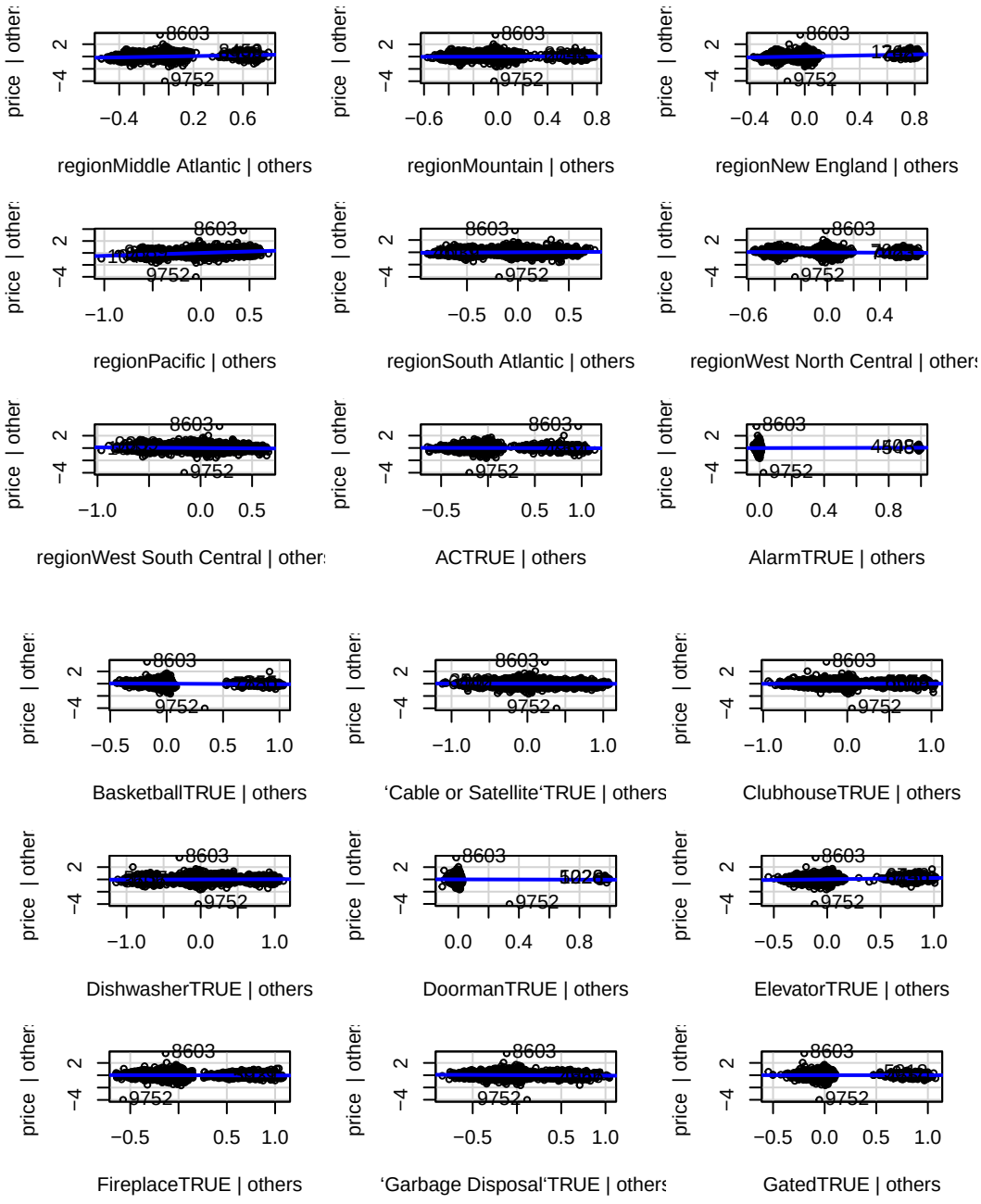

```
ggplot(data=rent_data_log) + geom_point(aes(x=bedrooms, y=price))
```

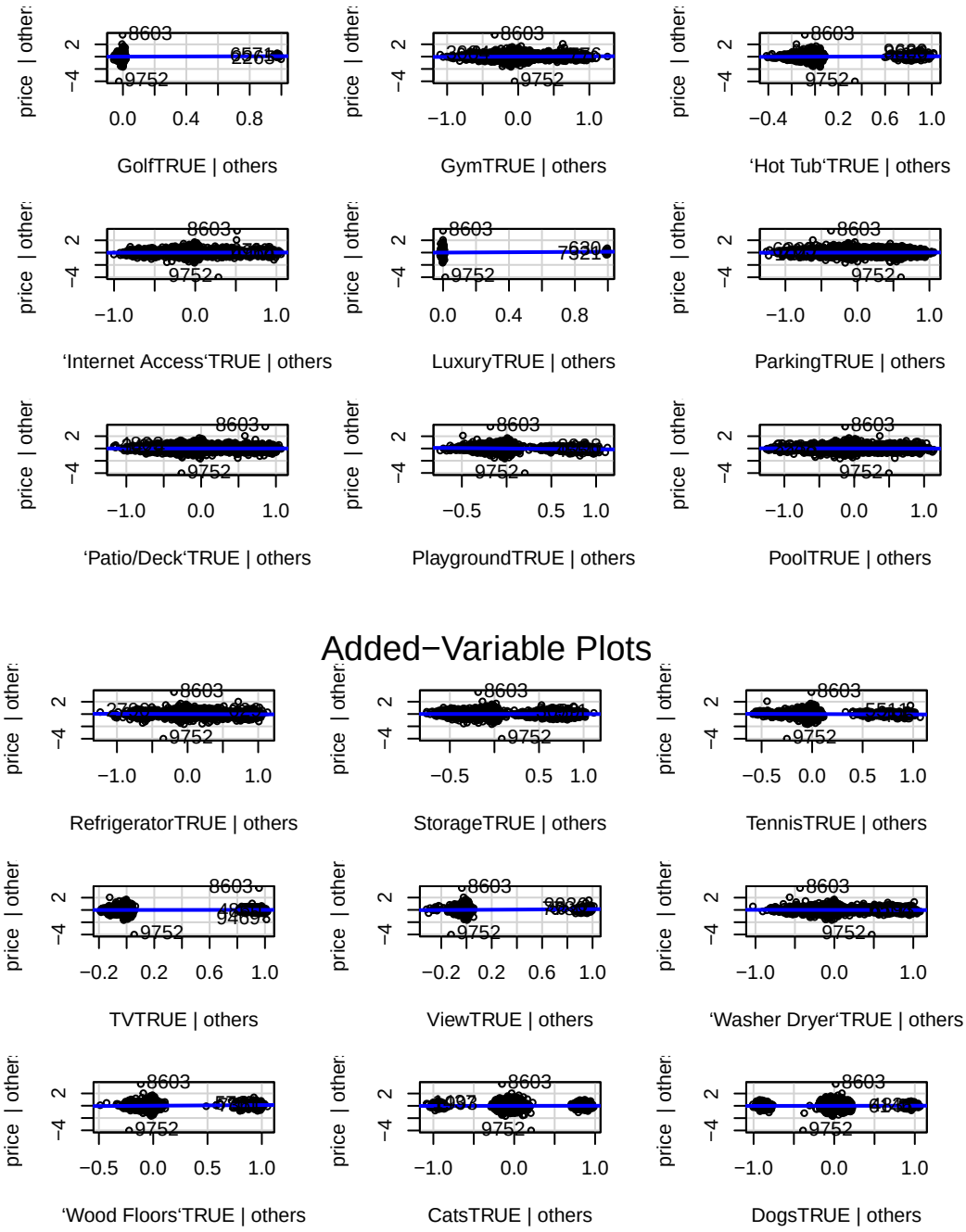


That one high point does still look influential after the transformation, but it is closer to the rest of the data points.

```
rent_log_lm %>% avPlots(ask = F)
```




price | others
regionMiddle Atlantic | others
regionMountain | others
regionNew England | others
regionPacific | others
regionSouth Atlantic | others
regionWest North Central | others
regionWest South Central | others
ACTRUE | others
AlarmTRUE | others
BasketballTRUE | others
'Cable or Satellite'TRUE | others
ClubhouseTRUE | others
DishwasherTRUE | others
DoormanTRUE | others
ElevatorTRUE | others
FireplaceTRUE | others
'Garbage Disposal'TRUE | others
GatedTRUE | others


Added−Variable Plots
price | others
GolfTRUE | others
GymTRUE | others
‘Hot Tub‘TRUE | others
‘Internet Access‘TRUE | others
LuxuryTRUE | others
ParkingTRUE | others
‘Patio/Deck‘TRUE | others
PlaygroundTRUE | others
PoolTRUE | others
RefrigeratorTRUE | others
StorageTRUE | others
TennisTRUE | others
TVTRUE | others
ViewTRUE | others
‘Washer Dryer‘TRUE | others
‘Wood Floors‘TRUE | others
CatsTRUE | others
DogsTRUE | others
8603
9752

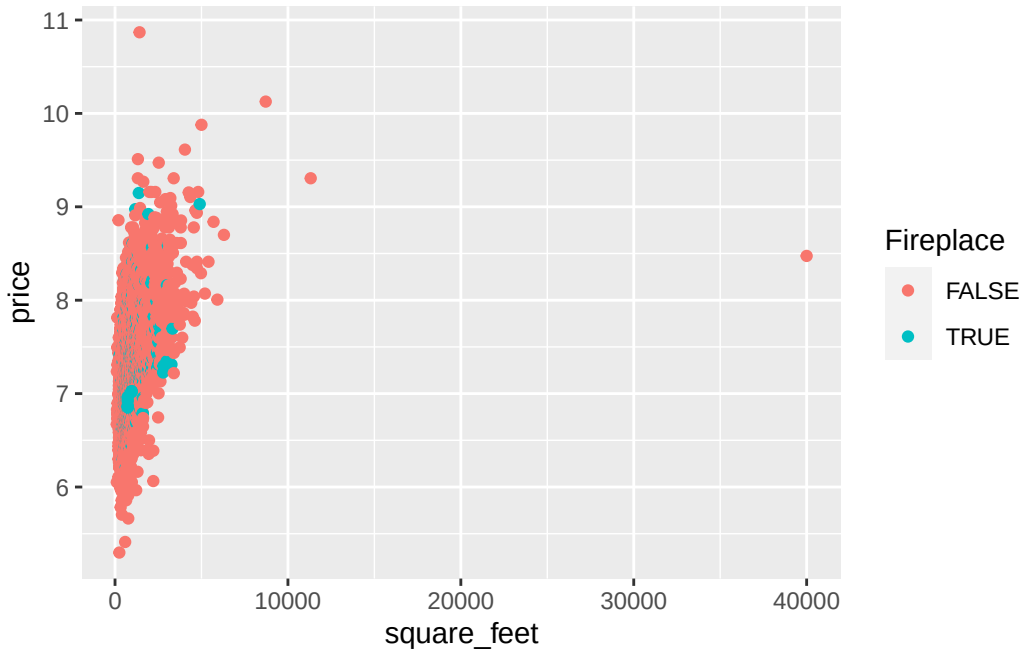
Variable selection

```r
rent_lm_reduced = rent_log_lm %>% step(
  direction = "both",
  scope = formula(rent_log_lm),
  trace = 0
)
summary(rent_lm_reduced)
```

```
Call:
lm(formula = price ~ category + bathrooms + bedrooms + has_photo +
    square_feet + agi + population + region + AC + Basketball +
    `Cable or Satellite` + Dishwasher + Elevator + Fireplace +
    `Garbage Disposal` + Gym + `Hot Tub` + `Internet Access` +
    Luxury + Parking + Playground + Pool + Refrigerator + Tennis +
    View + `Washer Dryer` + `Wood Floors` + Dogs, data = rent_data_log)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0214 -0.2015 -0.0139  0.1916  3.5616

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               6.534e+00  2.770e-02 235.871  < 2e-16 ***
categoryhome             -6.696e-01  2.251e-01  -2.975 0.002939 **
categoryshort_term       -5.574e-01  2.243e-01  -2.485 0.012974 *
bathrooms                 1.937e-01  8.019e-03  24.156  < 2e-16 ***
bedrooms                  5.710e-02  5.151e-03  11.087  < 2e-16 ***
has_photoTRUE            -3.882e-02  2.394e-02  -1.622 0.104940
square_feet               1.209e-04  6.498e-06  18.603  < 2e-16 ***
agi                       8.973e-08  3.174e-09  28.274  < 2e-16 ***
population               -1.445e-06  2.020e-07  -7.154 9.07e-13 ***
regionEast South Central -7.957e-02  2.445e-02  -3.254 0.001141 **
regionMiddle Atlantic     3.383e-01  1.589e-02  21.290  < 2e-16 ***
regionMountain            5.680e-02  1.543e-02   3.681 0.000234 ***
regionNew England         4.001e-01  1.915e-02  20.889  < 2e-16 ***
regionPacific             4.824e-01  1.248e-02  38.649  < 2e-16 ***
regionSouth Atlantic      8.016e-02  1.210e-02   6.624 3.67e-11 ***
regionWest North Central -9.826e-02  1.398e-02  -7.029 2.22e-12 ***
regionWest South Central -1.046e-01  1.222e-02  -8.559  < 2e-16 ***
```

```
ACTRUE                    -3.471e-02  1.494e-02  -2.323 0.020193 *
BasketballTRUE            -8.518e-02  1.969e-02  -4.326 1.54e-05 ***
`Cable or Satellite`TRUE -1.873e-02  1.255e-02  -1.492 0.135669
DishwasherTRUE             3.272e-02  1.037e-02   3.154 0.001613 **
ElevatorTRUE               2.113e-01  1.420e-02  14.878  < 2e-16 ***
FireplaceTRUE             -3.503e-02  1.132e-02  -3.094 0.001980 **
`Garbage Disposal`TRUE   -7.099e-02  1.235e-02  -5.746 9.39e-09 ***
GymTRUE                    4.737e-02  1.143e-02   4.145 3.42e-05 ***
`Hot Tub`TRUE              6.242e-02  1.859e-02   3.358 0.000788 ***
`Internet Access`TRUE     1.952e-02  1.228e-02   1.590 0.111767
LuxuryTRUE                 1.355e-01  9.582e-02   1.414 0.157446
ParkingTRUE                1.604e-02  7.868e-03   2.039 0.041449 *
PlaygroundTRUE            -1.235e-01  1.326e-02  -9.315  < 2e-16 ***
PoolTRUE                  -1.527e-02  8.719e-03  -1.752 0.079878 .
RefrigeratorTRUE          -3.598e-02  1.104e-02  -3.258 0.001127 **
TennisTRUE                -5.591e-02  1.662e-02  -3.365 0.000770 ***
ViewTRUE                   1.105e-01  2.684e-02   4.118 3.86e-05 ***
`Washer Dryer`TRUE       -1.835e-02  1.262e-02  -1.454 0.145886
`Wood Floors`TRUE          1.060e-01  1.774e-02   5.978 2.33e-09 ***
DogsTRUE                   2.307e-02  6.803e-03   3.391 0.000700 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3167 on 9724 degrees of freedom
Multiple R-squared:  0.545, Adjusted R-squared:  0.5433
F-statistic: 323.5 on 36 and 9724 DF,  p-value: < 2.2e-16
```

```
ggplot(data=rent_data_log) + geom_point(aes(x=square_feet, y=price, color=Fireplace))
```



It seems strange that fireplace is negative, so I decided to look at square footage since that might be an influence. There is one very large house. We'll remove that

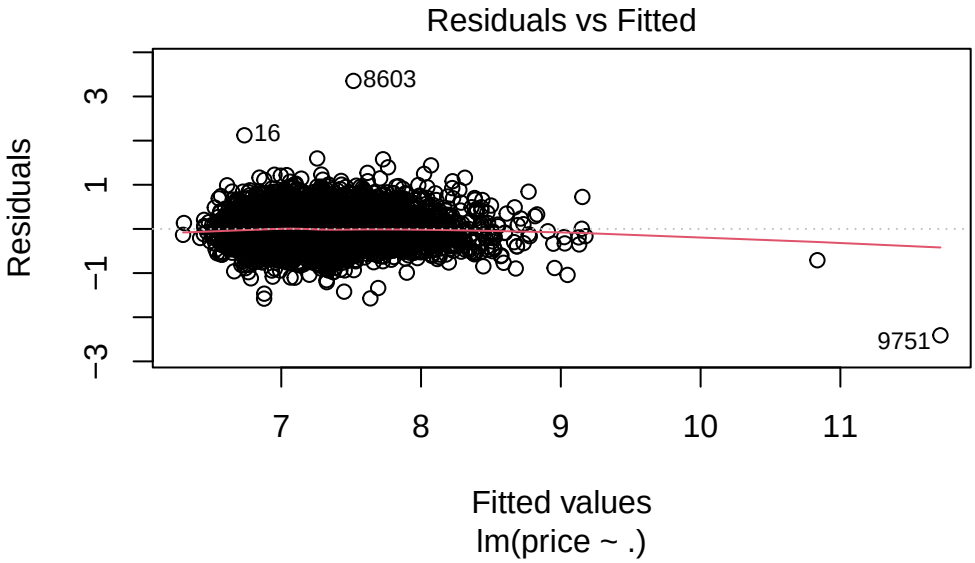
```
rent_data_log_reduced = rent_data_log %>%
  filter(square_feet <= 39000) %>%
  select(where(~ n_distinct(.) > 1))
rent_lm_log_reduced = lm(price ~ ., data=rent_data_log_reduced)
summary(rent_lm_log_reduced)
```

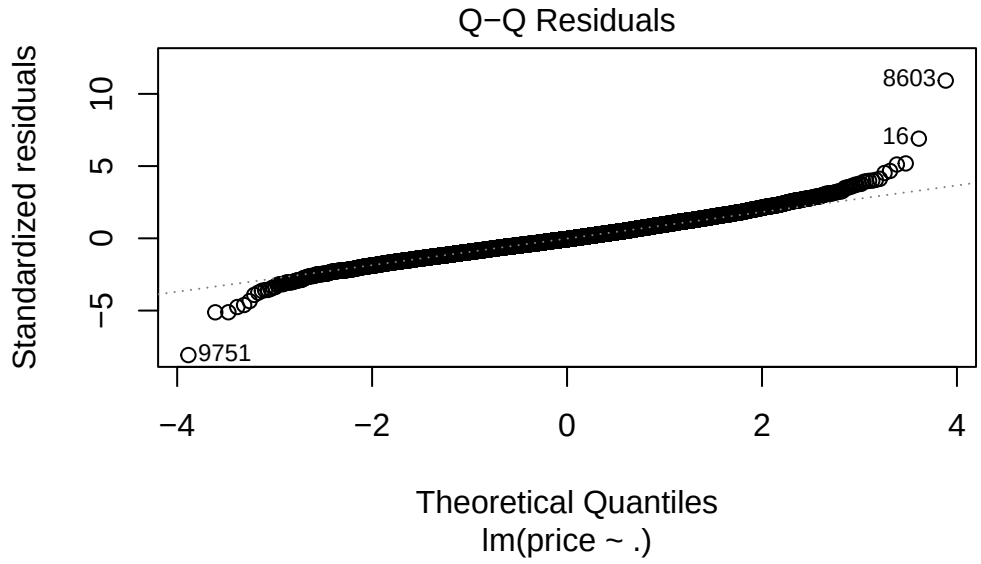
```
Call:
lm(formula = price ~ ., data = rent_data_log_reduced)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4099 -0.1961 -0.0170  0.1861  3.3522

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              6.531e+00  2.719e-02 240.188  < 2e-16 ***
categoryhome            -6.794e-01  2.196e-01  -3.093 0.001985 **
categoryshort_term      -5.770e-01  2.188e-01  -2.638 0.008364 **
```

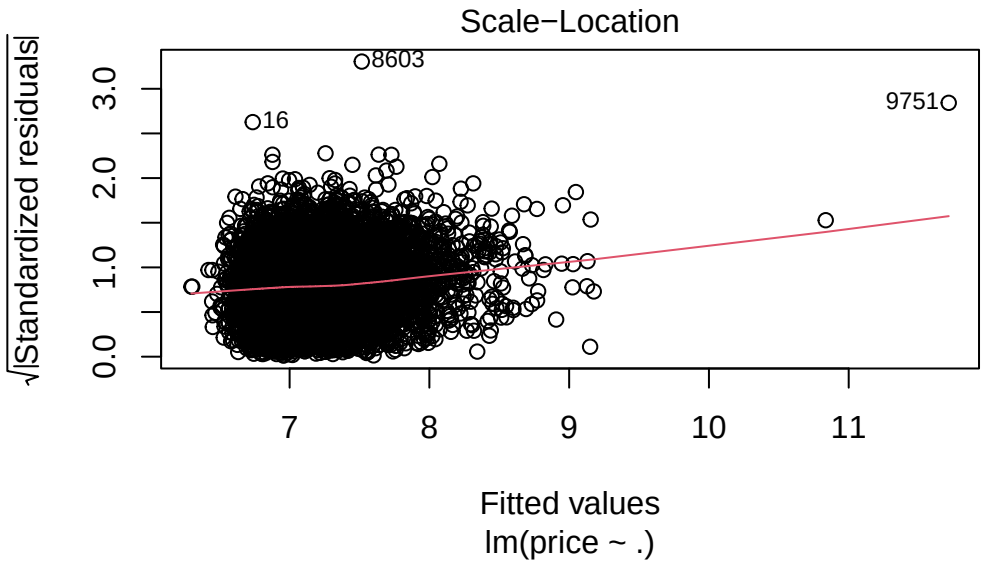
```
bathrooms                  1.008e-01  8.848e-03  11.395  < 2e-16 ***
bedrooms                   2.066e-02  5.288e-03   3.907 9.40e-05 ***
has_photoTRUE             -3.678e-02  2.336e-02  -1.574 0.115442
square_feet                3.172e-04  1.079e-05  29.394  < 2e-16 ***
agi                        9.274e-08  3.103e-09  29.888  < 2e-16 ***
population                -1.398e-06  1.976e-07  -7.078 1.56e-12 ***
regionEast South Central  -7.389e-02  2.390e-02  -3.091 0.001999 **
regionMiddle Atlantic      3.377e-01  1.554e-02  21.738  < 2e-16 ***
regionMountain             5.618e-02  1.516e-02   3.705 0.000212 ***
regionNew England          3.951e-01  1.875e-02  21.077  < 2e-16 ***
regionPacific              4.817e-01  1.227e-02  39.250  < 2e-16 ***
regionSouth Atlantic       7.724e-02  1.190e-02   6.492 8.88e-11 ***
regionWest North Central  -1.047e-01  1.365e-02  -7.668 1.91e-14 ***
regionWest South Central  -1.072e-01  1.209e-02  -8.868  < 2e-16 ***
ACTRUE                    -4.013e-02  1.468e-02  -2.735 0.006258 **
AlarmTRUE                  6.551e-02  6.488e-02   1.010 0.312684
BasketballTRUE            -7.038e-02  1.926e-02  -3.653 0.000260 ***
`Cable or Satellite`TRUE  -1.148e-02  1.240e-02  -0.925 0.354752
ClubhouseTRUE              8.014e-03  1.135e-02   0.706 0.479995
DishwasherTRUE             3.267e-02  1.021e-02   3.199 0.001381 **
DoormanTRUE                1.061e-01  6.062e-02   1.751 0.080007 .
ElevatorTRUE               2.093e-01  1.393e-02  15.021  < 2e-16 ***
FireplaceTRUE             -4.530e-02  1.126e-02  -4.024 5.76e-05 ***
`Garbage Disposal`TRUE    -7.012e-02  1.232e-02  -5.692 1.30e-08 ***
GatedTRUE                 -3.907e-03  1.570e-02  -0.249 0.803461
GolfTRUE                   6.518e-02  6.522e-02   0.999 0.317612
GymTRUE                    4.586e-02  1.142e-02   4.017 5.95e-05 ***
`Hot Tub`TRUE              7.606e-02  1.830e-02   4.155 3.28e-05 ***
`Internet Access`TRUE      2.281e-02  1.209e-02   1.887 0.059154 .
LuxuryTRUE                 1.508e-01  9.346e-02   1.614 0.106627
ParkingTRUE                2.235e-02  7.834e-03   2.854 0.004333 **
`Patio/Deck`TRUE          -9.768e-04  9.253e-03  -0.106 0.915933
PlaygroundTRUE            -1.198e-01  1.310e-02  -9.140  < 2e-16 ***
PoolTRUE                  -1.175e-02  8.639e-03  -1.360 0.173952
RefrigeratorTRUE          -4.215e-02  1.082e-02  -3.896 9.87e-05 ***
StorageTRUE               -7.758e-03  9.708e-03  -0.799 0.424259
TennisTRUE                -6.434e-02  1.625e-02  -3.960 7.54e-05 ***
TVTRUE                     1.750e-02  2.250e-02   0.778 0.436595
ViewTRUE                   1.027e-01  2.634e-02   3.901 9.65e-05 ***
`Washer Dryer`TRUE        -9.202e-03  1.241e-02  -0.742 0.458210
`Wood Floors`TRUE          9.913e-02  1.738e-02   5.705 1.20e-08 ***
CatsTRUE                   1.641e-02  1.371e-02   1.197 0.231294
DogsTRUE                   4.302e-03  1.354e-02   0.318 0.750652
```

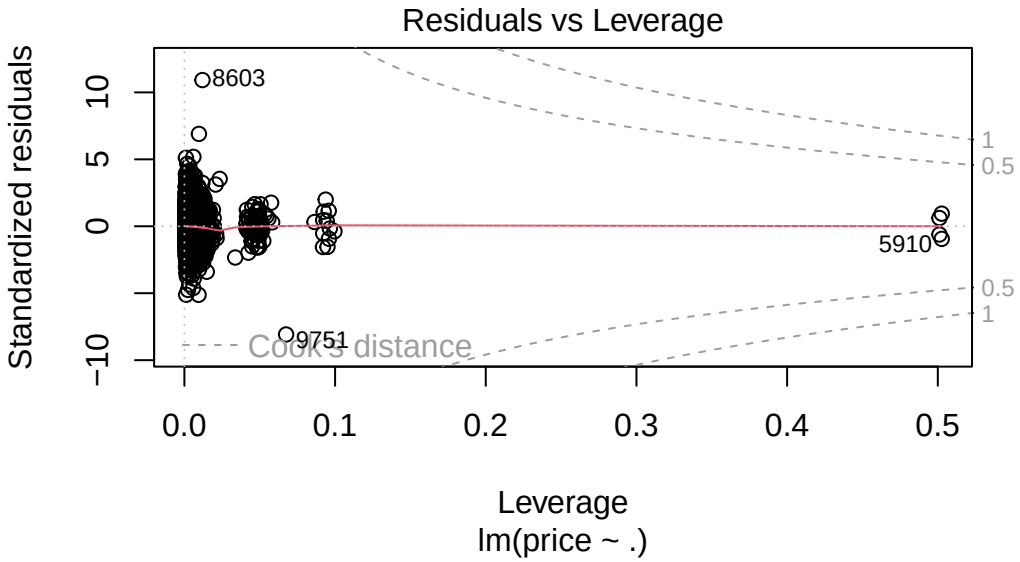
```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3088 on 9714 degrees of freedom
Multiple R-squared:  0.5674,    Adjusted R-squared:  0.5654
F-statistic: 283.1 on 45 and 9714 DF,  p-value: < 2.2e-16
```

```r
rent_lm_log_reduced %>% plot()
```










```
step_lm = rent_lm_log_reduced %>% step(
  direction = "both",
  scope = formula(rent_lm_log_reduced),
  trace = 0
)
summary(step_lm)
```

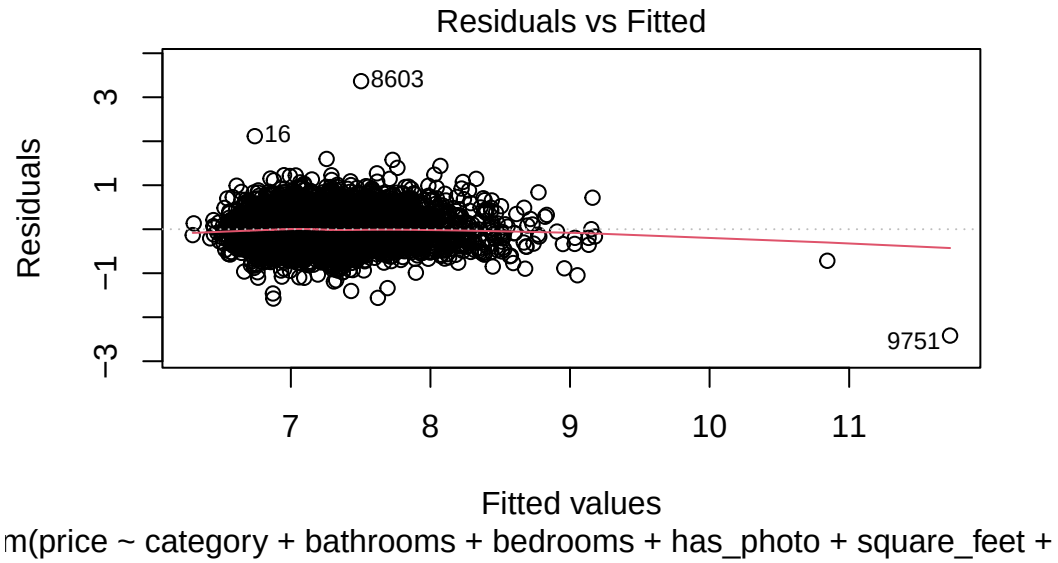
```
Call:
lm(formula = price ~ category + bathrooms + bedrooms + has_photo +
    square_feet + agi + population + region + AC + Basketball +
    Dishwasher + Doorman + Elevator + Fireplace + `Garbage Disposal` +
    Gym + `Hot Tub` + `Internet Access` + Luxury + Parking +
    Playground + Refrigerator + Tennis + View + `Wood Floors` +
    Cats, data = rent_data_log_reduced)

Residuals:
    Min      1Q  Median      3Q     Max
-2.4172 -0.1963 -0.0168  0.1857  3.3654

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               6.529e+00  2.707e-02 241.207  < 2e-16 ***
categoryhome             -6.894e-01  2.193e-01  -3.144 0.001674 **
categoryshort_term       -5.851e-01  2.186e-01  -2.676 0.007452 **
bathrooms                 1.010e-01  8.831e-03  11.434  < 2e-16 ***
bedrooms                  2.124e-02  5.267e-03   4.033 5.55e-05 ***
has_photoTRUE            -3.571e-02  2.330e-02  -1.532 0.125438
square_feet               3.178e-04  1.077e-05  29.511  < 2e-16 ***
agi                       9.260e-08  3.086e-09  30.010  < 2e-16 ***
population               -1.400e-06  1.972e-07  -7.100 1.33e-12 ***
regionEast South Central -7.434e-02  2.379e-02  -3.125 0.001784 **
regionMiddle Atlantic     3.365e-01  1.545e-02  21.784  < 2e-16 ***
regionMountain            5.550e-02  1.495e-02   3.713 0.000206 ***
regionNew England         3.943e-01  1.866e-02  21.133  < 2e-16 ***
regionPacific             4.816e-01  1.218e-02  39.552  < 2e-16 ***
regionSouth Atlantic      7.543e-02  1.164e-02   6.479 9.70e-11 ***
regionWest North Central -1.048e-01  1.361e-02  -7.704 1.45e-14 ***
regionWest South Central -1.090e-01  1.176e-02  -9.270  < 2e-16 ***
ACTRUE                   -4.360e-02  1.395e-02  -3.125 0.001786 **
BasketballTRUE           -7.140e-02  1.915e-02  -3.728 0.000194 ***
DishwasherTRUE            3.060e-02  1.002e-02   3.054 0.002262 **
DoormanTRUE               1.038e-01  6.044e-02   1.718 0.085866 .
ElevatorTRUE              2.084e-01  1.387e-02  15.028  < 2e-16 ***
FireplaceTRUE            -4.726e-02  1.080e-02  -4.377 1.22e-05 ***
`Garbage Disposal`TRUE   -7.115e-02  1.186e-02  -6.000 2.04e-09 ***
GymTRUE                   4.066e-02  1.014e-02   4.009 6.15e-05 ***
`Hot Tub`TRUE             7.349e-02  1.798e-02   4.087 4.41e-05 ***
`Internet Access`TRUE     1.882e-02  1.068e-02   1.762 0.078175 .
LuxuryTRUE                1.517e-01  9.342e-02   1.623 0.104531
ParkingTRUE               1.826e-02  7.500e-03   2.435 0.014913 *
```

```
PlaygroundTRUE           -1.229e-01  1.274e-02  -9.647  < 2e-16 ***
RefrigeratorTRUE         -4.337e-02  1.060e-02  -4.091 4.33e-05 ***
TennisTRUE               -6.696e-02  1.609e-02  -4.162 3.18e-05 ***
ViewTRUE                  1.003e-01  2.624e-02   3.824 0.000132 ***
`Wood Floors`TRUE         9.671e-02  1.715e-02   5.638 1.77e-08 ***
CatsTRUE                  1.938e-02  6.713e-03   2.887 0.003896 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3087 on 9725 degrees of freedom
Multiple R-squared:  0.5671,    Adjusted R-squared:  0.5656
F-statistic: 374.7 on 34 and 9725 DF,  p-value: < 2.2e-16
```

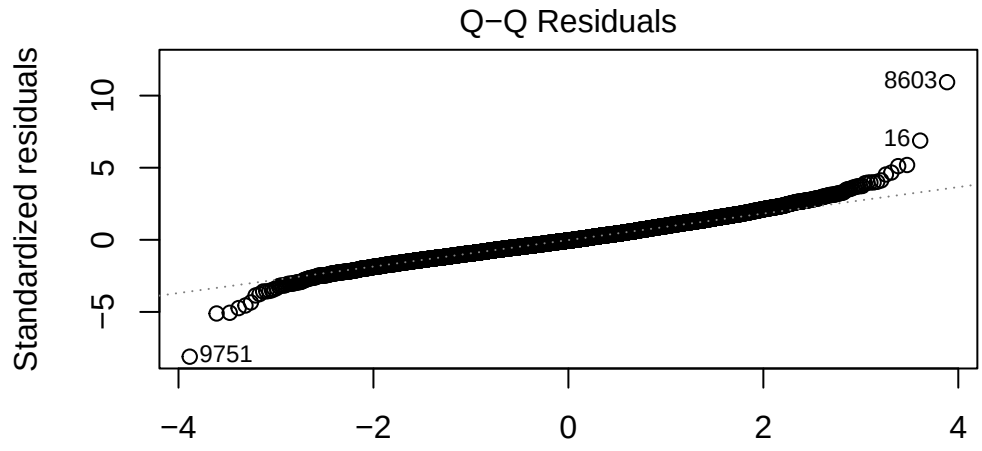
Interesting things:

- Dogs is not significant, but Cats is
- Which of the amenities were removed (Pool, Washer/Dryer, Clubhouse stood out)
- How some of the amentities (fireplace, refrigerator, tennis, etc) seem to bring the price *down* rather than up
  - Probably some relationship with the other amenities and/or aspects of a listing (square footage, etc) affecting the model

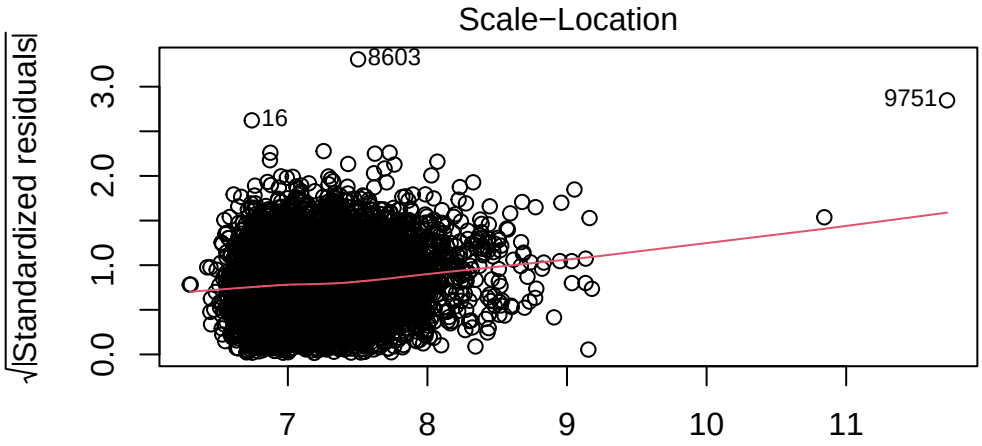
```
step_lm %>% plot()
```



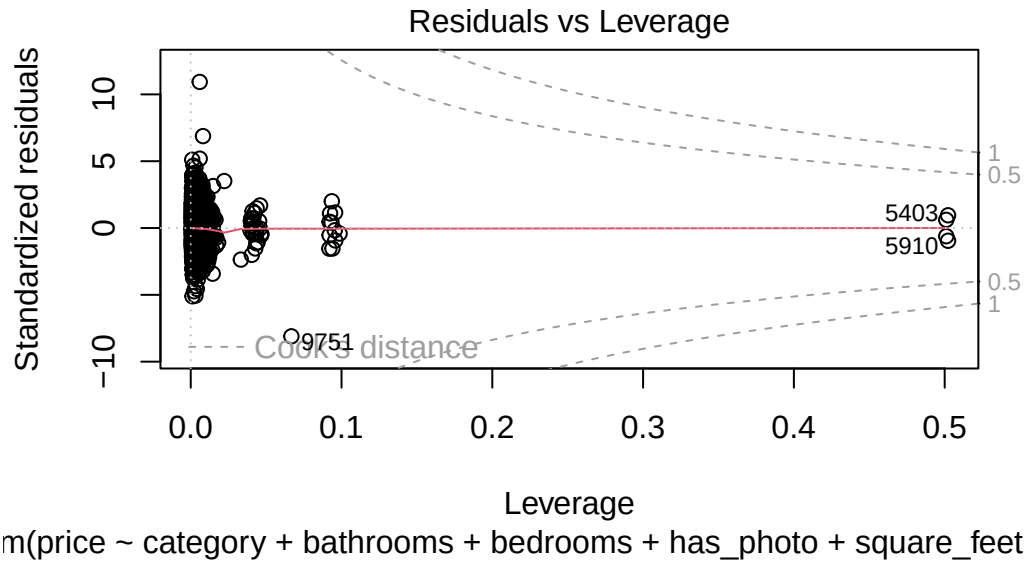
m(price ~ category + bathrooms + bedrooms + has_photo + square_feet +



m(price ~ category + bathrooms + bedrooms + has_photo + square_feet +



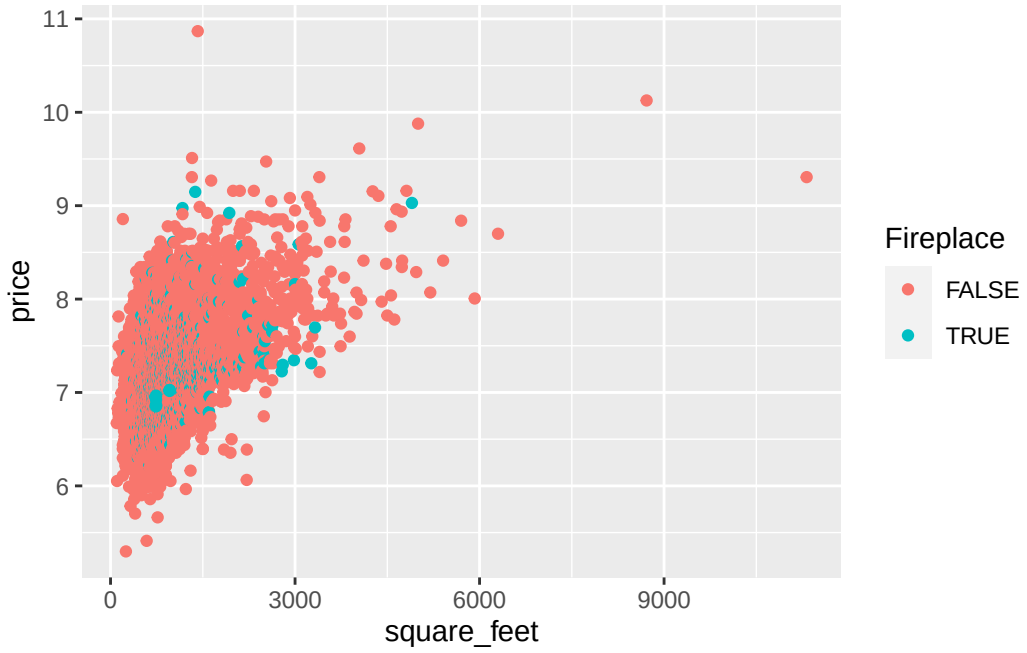
m(price ~ category + bathrooms + bedrooms + has_photo + square_feet +



m(price ~ category + bathrooms + bedrooms + has_photo + square_feet +

Some curving at the tails of the QQ-plot to take note of.

```
ggplot(data=rent_data_log_reduced) + geom_point(aes(x=square_feet, y=price, color=Fireplac
```



Still some outliers, probably making the "no fireplace" correlate with higher price/more square-footage

Quadratic/Cubic regression

```
rent_data_reduced = rent_data_log_reduced %>% select(
  price,
  category,
  bathrooms,
  bedrooms,
  has_photo,
  square_feet,
  agi,
  population,
  region,
  AC,
  Basketball,
  Dishwasher,
  Doorman,
  Elevator,
  Fireplace,
  `Garbage Disposal`,
  Gym,
  `Hot Tub`,
  `Internet Access`,
  Luxury,
  Parking,
  Playground,
  Refrigerator,
  Tennis,
  View,
  `Wood Floors`,
  Cats
) %>% mutate(
  bedrooms_sq = bedrooms^2,
  bathrooms_sq = bathrooms^2,
  square_feet_sq = square_feet^2,
  agi_sq = agi^2,
  population_sq = population^2,
)
quad_rent_lm = lm(price ~ ., data=rent_data_reduced)
summary(quad_rent_lm)
```

```
Call:
lm(formula = price ~ ., data = rent_data_reduced)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6062 -0.1982 -0.0178  0.1863  3.2652

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               6.525e+00  3.116e-02 209.404  < 2e-16 ***
categoryhome             -6.757e-01  2.169e-01  -3.115 0.001845 **
categoryshort_term       -6.134e-01  2.163e-01  -2.835 0.004589 **
bathrooms                 2.686e-02  2.224e-02   1.208 0.227191
bedrooms                 -1.612e-02  1.184e-02  -1.362 0.173153
has_photoTRUE            -3.219e-02  2.306e-02  -1.396 0.162750
square_feet               5.090e-04  1.777e-05  28.642  < 2e-16 ***
agi                       1.118e-07  8.477e-09  13.192  < 2e-16 ***
population               -3.214e-06  6.179e-07  -5.201 2.03e-07 ***
regionEast South Central -7.559e-02  2.354e-02  -3.211 0.001326 **
regionMiddle Atlantic     3.342e-01  1.530e-02  21.839  < 2e-16 ***
regionMountain            6.421e-02  1.492e-02   4.302 1.71e-05 ***
regionNew England         3.954e-01  1.853e-02  21.336  < 2e-16 ***
regionPacific             4.875e-01  1.217e-02  40.053  < 2e-16 ***
regionSouth Atlantic      7.542e-02  1.166e-02   6.471 1.02e-10 ***
regionWest North Central -1.039e-01  1.349e-02  -7.698 1.52e-14 ***
regionWest South Central -1.089e-01  1.173e-02  -9.290  < 2e-16 ***
ACTRUE                   -4.436e-02  1.382e-02  -3.210 0.001331 **
BasketballTRUE           -6.864e-02  1.894e-02  -3.624 0.000292 ***
DishwasherTRUE            2.976e-02  9.923e-03   2.999 0.002715 **
DoormanTRUE               1.114e-01  5.978e-02   1.864 0.062337 .
ElevatorTRUE              2.095e-01  1.372e-02  15.266  < 2e-16 ***
FireplaceTRUE            -5.163e-02  1.069e-02  -4.829 1.39e-06 ***
`Garbage Disposal`TRUE   -7.057e-02  1.173e-02  -6.017 1.84e-09 ***
GymTRUE                   3.705e-02  1.007e-02   3.681 0.000233 ***
`Hot Tub`TRUE             7.347e-02  1.781e-02   4.126 3.72e-05 ***
`Internet Access`TRUE     2.350e-02  1.058e-02   2.222 0.026311 *
LuxuryTRUE                1.570e-01  9.241e-02   1.700 0.089249 .
ParkingTRUE               1.868e-02  7.421e-03   2.517 0.011846 *
PlaygroundTRUE           -1.230e-01  1.261e-02  -9.758  < 2e-16 ***
RefrigeratorTRUE         -4.006e-02  1.056e-02  -3.795 0.000149 ***
TennisTRUE               -6.797e-02  1.592e-02  -4.271 1.97e-05 ***
ViewTRUE                  1.012e-01  2.595e-02   3.897 9.79e-05 ***
`Wood Floors`TRUE         9.345e-02  1.698e-02   5.505 3.79e-08 ***
```

```
CatsTRUE                  1.563e-02  6.659e-03   2.347 0.018937 *
bedrooms_sq               4.699e-03  2.396e-03   1.961 0.049862 *
bathrooms_sq              1.738e-02  5.242e-03   3.316 0.000916 ***
square_feet_sq           -4.595e-08  3.650e-09 -12.589  < 2e-16 ***
agi_sq                   -3.006e-15  1.476e-15  -2.036 0.041806 *
population_sq             1.827e-11  6.421e-12   2.846 0.004442 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3053 on 9720 degrees of freedom
Multiple R-squared:  0.5768,    Adjusted R-squared:  0.5751
F-statistic: 339.7 on 39 and 9720 DF,  p-value: < 2.2e-16
```

Looks like all of these factors may be quadratic

```
summary(quad_rent_lm %>% step(
  direction = "both",
  scope = formula(quad_rent_lm),
  trace = 0
))
```

```
Call:
lm(formula = price ~ category + square_feet + agi + population +
    region + AC + Basketball + Dishwasher + Doorman + Elevator +
    Fireplace + `Garbage Disposal` + Gym + `Hot Tub` + `Internet Access` +
    Luxury + Parking + Playground + Refrigerator + Tennis + View +
    `Wood Floors` + Cats + bedrooms_sq + bathrooms_sq + square_feet_sq +
    agi_sq + population_sq, data = rent_data_reduced)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6043 -0.1981 -0.0168  0.1865  3.2761

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              6.495e+00  1.709e-02 380.175  < 2e-16 ***
categoryhome            -6.793e-01  2.169e-01  -3.132 0.001744 **
categoryshort_term      -6.272e-01  2.162e-01  -2.902 0.003721 **
square_feet              5.162e-04  1.426e-05  36.186  < 2e-16 ***
agi                      1.122e-07  8.474e-09  13.240  < 2e-16 ***
```

```
population                -3.204e-06  6.178e-07  -5.186 2.19e-07 ***
regionEast South Central  -7.448e-02  2.352e-02  -3.166 0.001548 **
regionMiddle Atlantic      3.354e-01  1.528e-02  21.947  < 2e-16 ***
regionMountain             6.519e-02  1.491e-02   4.372 1.25e-05 ***
regionNew England          3.944e-01  1.851e-02  21.304  < 2e-16 ***
regionPacific              4.888e-01  1.215e-02  40.216  < 2e-16 ***
regionSouth Atlantic       7.660e-02  1.164e-02   6.583 4.84e-11 ***
regionWest North Central  -1.046e-01  1.349e-02  -7.753 9.87e-15 ***
regionWest South Central  -1.080e-01  1.172e-02  -9.219  < 2e-16 ***
ACTRUE                    -4.456e-02  1.381e-02  -3.227 0.001256 **
BasketballTRUE            -6.853e-02  1.894e-02  -3.618 0.000298 ***
DishwasherTRUE             2.888e-02  9.912e-03   2.914 0.003578 **
DoormanTRUE                1.133e-01  5.977e-02   1.896 0.057985 .
ElevatorTRUE               2.098e-01  1.371e-02  15.296  < 2e-16 ***
FireplaceTRUE             -5.087e-02  1.069e-02  -4.761 1.96e-06 ***
`Garbage Disposal`TRUE    -7.053e-02  1.173e-02  -6.013 1.88e-09 ***
GymTRUE                    3.759e-02  1.004e-02   3.743 0.000183 ***
`Hot Tub`TRUE              7.334e-02  1.780e-02   4.119 3.84e-05 ***
`Internet Access`TRUE      2.351e-02  1.058e-02   2.222 0.026283 *
LuxuryTRUE                 1.599e-01  9.231e-02   1.732 0.083321 .
ParkingTRUE                1.890e-02  7.419e-03   2.548 0.010858 *
PlaygroundTRUE            -1.236e-01  1.261e-02  -9.807  < 2e-16 ***
RefrigeratorTRUE          -3.970e-02  1.055e-02  -3.762 0.000170 ***
TennisTRUE                -6.792e-02  1.592e-02  -4.267 2.00e-05 ***
ViewTRUE                   1.012e-01  2.595e-02   3.899 9.72e-05 ***
`Wood Floors`TRUE          9.388e-02  1.697e-02   5.531 3.27e-08 ***
CatsTRUE                   1.511e-02  6.653e-03   2.271 0.023151 *
bedrooms_sq                1.807e-03  1.089e-03   1.659 0.097227 .
bathrooms_sq               2.301e-02  2.066e-03  11.139  < 2e-16 ***
square_feet_sq            -4.779e-08  2.625e-09 -18.206  < 2e-16 ***
agi_sq                    -3.049e-15  1.476e-15  -2.065 0.038906 *
population_sq              1.816e-11  6.420e-12   2.829 0.004680 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3054 on 9723 degrees of freedom
Multiple R-squared:  0.5766,    Adjusted R-squared:  0.575
F-statistic: 367.8 on 36 and 9723 DF,  p-value: < 2.2e-16
```

All of the previously-identified variables are still relevant

```r
summary(quad_rent_lm %>% step(
  direction = "forward",
  scope = formula(quad_rent_lm),
  trace = 1
))
```

```
Start:  AIC=-23116.95
price ~ category + bathrooms + bedrooms + has_photo + square_feet +
    agi + population + region + AC + Basketball + Dishwasher +
    Doorman + Elevator + Fireplace + `Garbage Disposal` + Gym +
    `Hot Tub` + `Internet Access` + Luxury + Parking + Playground +
    Refrigerator + Tennis + View + `Wood Floors` + Cats + bedrooms_sq +
    bathrooms_sq + square_feet_sq + agi_sq + population_sq



Call:
lm(formula = price ~ category + bathrooms + bedrooms + has_photo +
    square_feet + agi + population + region + AC + Basketball +
    Dishwasher + Doorman + Elevator + Fireplace + `Garbage Disposal` +
    Gym + `Hot Tub` + `Internet Access` + Luxury + Parking +
    Playground + Refrigerator + Tennis + View + `Wood Floors` +
    Cats + bedrooms_sq + bathrooms_sq + square_feet_sq + agi_sq +
    population_sq, data = rent_data_reduced)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6062 -0.1982 -0.0178  0.1863  3.2652

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               6.525e+00  3.116e-02 209.404  < 2e-16 ***
categoryhome             -6.757e-01  2.169e-01  -3.115 0.001845 **
categoryshort_term       -6.134e-01  2.163e-01  -2.835 0.004589 **
bathrooms                 2.686e-02  2.224e-02   1.208 0.227191
bedrooms                 -1.612e-02  1.184e-02  -1.362 0.173153
has_photoTRUE            -3.219e-02  2.306e-02  -1.396 0.162750
square_feet               5.090e-04  1.777e-05  28.642  < 2e-16 ***
agi                       1.118e-07  8.477e-09  13.192  < 2e-16 ***
population               -3.214e-06  6.179e-07  -5.201 2.03e-07 ***
regionEast South Central -7.559e-02  2.354e-02  -3.211 0.001326 **
regionMiddle Atlantic     3.342e-01  1.530e-02  21.839  < 2e-16 ***
```

```
regionMountain             6.421e-02  1.492e-02   4.302 1.71e-05 ***
regionNew England          3.954e-01  1.853e-02  21.336  < 2e-16 ***
regionPacific              4.875e-01  1.217e-02  40.053  < 2e-16 ***
regionSouth Atlantic       7.542e-02  1.166e-02   6.471 1.02e-10 ***
regionWest North Central -1.039e-01  1.349e-02  -7.698 1.52e-14 ***
regionWest South Central -1.089e-01  1.173e-02  -9.290  < 2e-16 ***
ACTRUE                   -4.436e-02  1.382e-02  -3.210 0.001331 **
BasketballTRUE           -6.864e-02  1.894e-02  -3.624 0.000292 ***
DishwasherTRUE            2.976e-02  9.923e-03   2.999 0.002715 **
DoormanTRUE               1.114e-01  5.978e-02   1.864 0.062337 .
ElevatorTRUE              2.095e-01  1.372e-02  15.266  < 2e-16 ***
FireplaceTRUE            -5.163e-02  1.069e-02  -4.829 1.39e-06 ***
`Garbage Disposal`TRUE   -7.057e-02  1.173e-02  -6.017 1.84e-09 ***
GymTRUE                   3.705e-02  1.007e-02   3.681 0.000233 ***
`Hot Tub`TRUE             7.347e-02  1.781e-02   4.126 3.72e-05 ***
`Internet Access`TRUE     2.350e-02  1.058e-02   2.222 0.026311 *
LuxuryTRUE                1.570e-01  9.241e-02   1.700 0.089249 .
ParkingTRUE               1.868e-02  7.421e-03   2.517 0.011846 *
PlaygroundTRUE           -1.230e-01  1.261e-02  -9.758  < 2e-16 ***
RefrigeratorTRUE         -4.006e-02  1.056e-02  -3.795 0.000149 ***
TennisTRUE               -6.797e-02  1.592e-02  -4.271 1.97e-05 ***
ViewTRUE                  1.012e-01  2.595e-02   3.897 9.79e-05 ***
`Wood Floors`TRUE         9.345e-02  1.698e-02   5.505 3.79e-08 ***
CatsTRUE                  1.563e-02  6.659e-03   2.347 0.018937 *
bedrooms_sq               4.699e-03  2.396e-03   1.961 0.049862 *
bathrooms_sq              1.738e-02  5.242e-03   3.316 0.000916 ***
square_feet_sq           -4.595e-08  3.650e-09 -12.589  < 2e-16 ***
agi_sq                   -3.006e-15  1.476e-15  -2.036 0.041806 *
population_sq             1.827e-11  6.421e-12   2.846 0.004442 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3053 on 9720 degrees of freedom
Multiple R-squared:  0.5768,    Adjusted R-squared:  0.5751
F-statistic: 339.7 on 39 and 9720 DF,  p-value: < 2.2e-16
```

Looking at possible cubic terms as well. Square footage was the one that was still significant.

```
rent_data_reduced = rent_data_reduced %>% mutate(
  square_feet_cu = square_feet^3
)
cu_rent_lm = lm(price ~ ., data=rent_data_reduced)
```

```r
summary(cu_rent_lm)
```

```
Call:
lm(formula = price ~ ., data = rent_data_reduced)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6024 -0.1986 -0.0185  0.1842  3.2048

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                6.471e+00  3.175e-02 203.779  < 2e-16 ***
categoryhome              -6.762e-01  2.162e-01  -3.128 0.001767 **
categoryshort_term        -6.210e-01  2.156e-01  -2.880 0.003980 **
bathrooms                  1.701e-02  2.220e-02   0.767 0.443375
bedrooms                  -4.905e-02  1.246e-02  -3.936 8.35e-05 ***
has_photoTRUE             -2.714e-02  2.299e-02  -1.181 0.237780
square_feet                6.811e-04  2.747e-05  24.797  < 2e-16 ***
agi                        1.159e-07  8.463e-09  13.698  < 2e-16 ***
population                -3.287e-06  6.159e-07  -5.337 9.68e-08 ***
regionEast South Central  -7.544e-02  2.346e-02  -3.216 0.001305 **
regionMiddle Atlantic      3.320e-01  1.525e-02  21.770  < 2e-16 ***
regionMountain             7.097e-02  1.490e-02   4.765 1.92e-06 ***
regionNew England          3.955e-01  1.847e-02  21.414  < 2e-16 ***
regionPacific              4.923e-01  1.214e-02  40.537  < 2e-16 ***
regionSouth Atlantic       7.512e-02  1.162e-02   6.467 1.05e-10 ***
regionWest North Central  -1.025e-01  1.345e-02  -7.619 2.79e-14 ***
regionWest South Central  -1.073e-01  1.169e-02  -9.184  < 2e-16 ***
ACTRUE                    -4.505e-02  1.377e-02  -3.271 0.001074 **
BasketballTRUE            -6.770e-02  1.888e-02  -3.586 0.000338 ***
DishwasherTRUE             2.849e-02  9.891e-03   2.880 0.003985 **
DoormanTRUE                1.180e-01  5.958e-02   1.980 0.047732 *
ElevatorTRUE               2.113e-01  1.368e-02  15.453  < 2e-16 ***
FireplaceTRUE             -5.440e-02  1.066e-02  -5.103 3.40e-07 ***
`Garbage Disposal`TRUE    -7.116e-02  1.169e-02  -6.088 1.18e-09 ***
GymTRUE                    3.272e-02  1.005e-02   3.258 0.001128 **
`Hot Tub`TRUE              7.163e-02  1.775e-02   4.036 5.48e-05 ***
`Internet Access`TRUE      2.594e-02  1.055e-02   2.460 0.013910 *
LuxuryTRUE                 1.569e-01  9.209e-02   1.704 0.088432 .
ParkingTRUE                1.726e-02  7.398e-03   2.334 0.019628 *
```

```
PlaygroundTRUE           -1.240e-01  1.257e-02  -9.869  < 2e-16 ***
RefrigeratorTRUE         -3.821e-02  1.052e-02  -3.631 0.000284 ***
TennisTRUE               -6.815e-02  1.586e-02  -4.296 1.75e-05 ***
ViewTRUE                  1.010e-01  2.587e-02   3.906 9.43e-05 ***
`Wood Floors`TRUE         9.233e-02  1.692e-02   5.457 4.95e-08 ***
CatsTRUE                  1.293e-02  6.645e-03   1.945 0.051767 .
bedrooms_sq               1.177e-02  2.539e-03   4.638 3.57e-06 ***
bathrooms_sq              1.917e-02  5.229e-03   3.666 0.000248 ***
square_feet_sq           -1.207e-07  9.819e-09 -12.294  < 2e-16 ***
agi_sq                   -3.791e-15  1.475e-15  -2.571 0.010158 *
population_sq             1.865e-11  6.399e-12   2.915 0.003562 **
square_feet_cu            6.172e-12  7.529e-13   8.198 2.75e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3043 on 9719 degrees of freedom
Multiple R-squared:  0.5797,    Adjusted R-squared:  0.578
F-statistic: 335.1 on 40 and 9719 DF,  p-value: < 2.2e-16
```

```r
summary(cu_rent_lm %>% step(
  direction = "both",
  scope = formula(cu_rent_lm),
  trace = 0
))
```

```
Call:
lm(formula = price ~ category + bedrooms + square_feet + agi +
    population + region + AC + Basketball + Dishwasher + Doorman +
    Elevator + Fireplace + `Garbage Disposal` + Gym + `Hot Tub` +
    `Internet Access` + Luxury + Parking + Playground + Refrigerator +
    Tennis + View + `Wood Floors` + Cats + bedrooms_sq + bathrooms_sq +
    square_feet_sq + agi_sq + population_sq + square_feet_cu,
    data = rent_data_reduced)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6028 -0.1983 -0.0181  0.1839  3.2027

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)                6.450e+00  1.998e-02 322.846  < 2e-16 ***
categoryhome              -6.787e-01  2.162e-01  -3.139 0.001698 **
categoryshort_term        -6.283e-01  2.154e-01  -2.917 0.003546 **
bedrooms                  -4.614e-02  1.190e-02  -3.877 0.000106 ***
square_feet                6.900e-04  2.530e-05  27.270  < 2e-16 ***
agi                        1.161e-07  8.462e-09  13.716  < 2e-16 ***
population                -3.277e-06  6.158e-07  -5.322 1.05e-07 ***
regionEast South Central -7.484e-02  2.344e-02  -3.193 0.001413 **
regionMiddle Atlantic      3.329e-01  1.523e-02  21.855  < 2e-16 ***
regionMountain             7.166e-02  1.488e-02   4.815 1.49e-06 ***
regionNew England          3.951e-01  1.845e-02  21.413  < 2e-16 ***
regionPacific              4.932e-01  1.212e-02  40.675  < 2e-16 ***
regionSouth Atlantic       7.594e-02  1.160e-02   6.549 6.08e-11 ***
regionWest North Central -1.028e-01  1.344e-02  -7.645 2.29e-14 ***
regionWest South Central -1.068e-01  1.168e-02  -9.148  < 2e-16 ***
ACTRUE                    -4.488e-02  1.377e-02  -3.260 0.001117 **
BasketballTRUE            -6.751e-02  1.888e-02  -3.577 0.000350 ***
DishwasherTRUE             2.795e-02  9.882e-03   2.829 0.004681 **
DoormanTRUE                1.190e-01  5.957e-02   1.997 0.045825 *
ElevatorTRUE               2.113e-01  1.367e-02  15.451  < 2e-16 ***
FireplaceTRUE             -5.391e-02  1.065e-02  -5.060 4.27e-07 ***
`Garbage Disposal`TRUE    -7.108e-02  1.169e-02  -6.081 1.24e-09 ***
GymTRUE                    3.333e-02  1.003e-02   3.323 0.000893 ***
`Hot Tub`TRUE              7.135e-02  1.774e-02   4.021 5.84e-05 ***
`Internet Access`TRUE      2.592e-02  1.055e-02   2.458 0.013980 *
LuxuryTRUE                 1.608e-01  9.201e-02   1.748 0.080519 .
ParkingTRUE                1.744e-02  7.396e-03   2.358 0.018389 *
PlaygroundTRUE            -1.244e-01  1.256e-02  -9.902  < 2e-16 ***
RefrigeratorTRUE          -3.810e-02  1.052e-02  -3.621 0.000295 ***
TennisTRUE                -6.815e-02  1.586e-02  -4.297 1.75e-05 ***
ViewTRUE                   1.010e-01  2.586e-02   3.907 9.41e-05 ***
`Wood Floors`TRUE          9.273e-02  1.692e-02   5.482 4.31e-08 ***
CatsTRUE                   1.251e-02  6.638e-03   1.884 0.059544 .
bedrooms_sq                1.127e-02  2.458e-03   4.586 4.57e-06 ***
bathrooms_sq               2.282e-02  2.065e-03  11.052  < 2e-16 ***
square_feet_sq            -1.232e-07  9.404e-09 -13.098  < 2e-16 ***
agi_sq                    -3.826e-15  1.474e-15  -2.595 0.009467 **
population_sq              1.858e-11  6.398e-12   2.904 0.003697 **
square_feet_cu             6.227e-12  7.515e-13   8.286  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3043 on 9721 degrees of freedom
```

```
Multiple R-squared:  0.5796,    Adjusted R-squared:  0.578
F-statistic: 352.7 on 38 and 9721 DF,  p-value: < 2.2e-16
```

```
vif(cu_rent_lm)
```

```
                     GVIF Df GVIF^(1/(2*Df))
category             1.013052  2        1.003247
bathrooms           20.023260  1        4.474736
bedrooms            14.573301  1        3.817499
has_photo            1.019412  1        1.009659
square_feet         22.194214  1        4.711074
agi                 10.818831  1        3.289199
population          12.348284  1        3.514012
region               1.842687  8        1.038941
AC                   1.245922  1        1.116209
Basketball           1.169729  1        1.081540
Dishwasher           2.279664  1        1.509856
Doorman              1.032095  1        1.015921
Elevator             1.195562  1        1.093417
Fireplace            1.158605  1        1.076385
`Garbage Disposal`   1.545010  1        1.242984
Gym                  1.354392  1        1.163784
`Hot Tub`            1.116035  1        1.056425
`Internet Access`    1.459888  1        1.208258
Luxury               1.006271  1        1.003130
Parking              1.355417  1        1.164224
Playground           1.215366  1        1.102436
Refrigerator         2.523811  1        1.588651
Tennis               1.227733  1        1.108031
View                 1.053031  1        1.026173
`Wood Floors`        1.063230  1        1.031130
Cats                 1.138740  1        1.067117
bedrooms_sq         14.243272  1        3.774026
bathrooms_sq        20.181115  1        4.492340
square_feet_sq      55.775808  1        7.468320
agi_sq               9.264957  1        3.043839
population_sq       11.001572  1        3.316862
square_feet_cu      18.678266  1        4.321836
```

We have some more issues with multicollinearity, it appears

Summary

The following variables are impactful

- Category (apartment/house/short-term)
- Bedrooms (quadratic)
- Bathrooms (quadratic)
- Square footage (cubic)
- Income (quadratic)
- Population (quadratic)
- Region
- If cats were allowed
- The following amenities: AC, Basketball, Dishwasher, Doorman, Elevator, Fireplace, Garbage disposal, Gym, Hot tub, Internet, Luxury, Parking, Playground, Fridge, Tennis, View, Wood floors

Some surprising variables, like AC, fireplace, tennis, and basketball seem to have a negative relationship. If I had more time, I would explore all of these variables more to see if there is some multicollinearity that hasn't been caught, or some other explanation for these trends. I would also be interested in exploring the luxury variable more since I would expect it to have more of an impact, but maybe there is some correlation with the other amentities and luxury, depending on how "luxury" is defined

It also appears that population, AGI, number of bedrooms, square footage, and region have a large impact on price, in addition to a subset of the amentities included in the dataset.

Overall, while taking the log of price helped a lot with some of our assumptions/conditions, there still seems to be some problems that warrant further investigation. If I also had more time, I may look into using some other regression methods like Ridge or Lasso.