# (1) Data Cleaning: Distance to Food and Housing Prices

boseongyun

2024-10-10

## Contents

## Description of This File

This file outlines the data cleaning process in five primary steps:

1. **Sample Restriction**: We clean the data based on assumptions about the target population, requiring minimal assumptions about the data itself.
2. **Data Processing**: We refine data by understanding the data-generating process for key variables and modifying variable types as needed.
3. **Exploratory Data Analysis (EDA)**: We perform EDA to check that the data aligns with expectations, making adjustments where necessary.
4. **Final Assessment**: We evaluate the dataset to quantify uncertainties and ensure data quality.
5. **Data Saving**: We save the cleaned and processed dataset for analysis.

Below, we provide (1) a Project Description and (2) a Data Description to guide our data cleaning steps clearly.

## Project Description

This project examines whether access to food, measured by proximity to grocery stores, impacts housing prices. We focus specifically on single family households in Connecticut to explore the relationship between housing prices and access to food resources.

## Data Description

This analysis uses two primary datasets:

1. Housing data for Connecticut (in POLYGON format) and
2. A directory of stores authorized to accept SNAP benefits, with location coordinates (longitude and latitude).

The datasets are sourced from ATLAS. The housing data is spatially joined with store locations using the `st_nearest()` function, which calculates the nearest store to each housing location based on their geographic coordinates.

# Data Pre-Processing

## Import necessary Libraries and Data

```
library(sf)
```

```
## Linking to GEOS 3.11.0, GDAL 3.5.3, PROJ 9.1.0; sf_use_s2() is TRUE
```

Load the dataset and drop unnecessary geometry for efficiency

```
# Load the Connecticut Property Shapefile
# d <- st_read("CT-parcel-data/4c5501b8-b68e-4888-bf6a-d92670d69c3b.gdb/")

# For Easy Data Handling
# d <- st_transform(d, crs = 4326)
# h <- st_drop_geometry(d)  # Drop geometry data since it's not needed
# rm(d) # remove for efficiency
# names(h) <- tolower(names(h))  # Convert column names to lowercase
# saveRDS("ct.rds")
d <- readRDS("ct.rds")
```

## Main Step (1) Sample Restricton

### Data Filtering: Single-Family Homes in New Haven

We focus on single-family properties located in New Haven, CT. We refine the dataset by excluding:

1. Extremely expensive properties (outliers that may skew analysis)
2. Properties with a high number of rooms (likely multi-family/commercial)
3. Properties without essential components like bedroom or bathrooms

These filters help reduce noise and exclude outliers. We avoid dropping NA values initially, as they may provide valuable information later.

### Target Population: New Haven

```
# Filter for New Haven properties or missing town names (for inspection)
d <- d[d$town_name == "NEW HAVEN" | is.na(d$town_name), ]
nrow(d)  # Display the row count after filtering to verify reduction
```

```
## [1] 27378
```

### Filtering by Number of Bedrooms and Bathrooms

We keep properties with fewer than 5 bedrooms. Properties with 0 bedrooms or bathrooms are inspected for data accuracy. Cross-checked entries show that some NA values were incorrectly converted to 0, so these entries are removed.

For more information, we have verified through Cross-checking with Clear Vision, that many observations with 0 number of bedrooms and bathrooms have NA values. They have been converted to 0.

Since (1) we want the number of bedrooms and bathrooms to be greater than 0 and (2) because the variables have non-trivial problems. We have removed them from our data.

```
# Filter properties with fewer than 5 bedrooms
nrow(d)
```

```
## [1] 27378
```

```
sum(d$number_of_bedroom > 5, na.rm = TRUE)
```

```
## [1] 4417
```

```
d <- d[d$number_of_bedroom < 5 | is.na(d$number_of_bedroom), ]
head(d)  # Verify the dataset after filtering
```

```
##        town_name                link                    owner
## 754264 NEW HAVEN 52070-013 0853 00500 CITY OF NEW HAVEN AIRPORT
## 754265 NEW HAVEN 52070-014 0853 00100  PEREZ-RAMIREZ NOE MARTIN
## 754266 NEW HAVEN 52070-014 0853 00101 RODRIGUEZ WILLIAM & LYSIE
## 754267 NEW HAVEN 52070-014 0853 00200             GUEST CRAIG C
## 754268 NEW HAVEN 52070-014 0853 00300  MEADOWS BERNADETTE (EST)
## 754270 NEW HAVEN 52070-014 0853 00400        KILFEATHER FRANCIS
##                 co_owner          location   mailing_address mailing_city
## 754264 CITY OF NEW HAVEN   75 SOUTH END RD     165 CHURCH ST    NEW HAVEN
## 754265              <NA>  199 SOUTH END RD  199 SOUTH END RD    NEW HAVEN
## 754266              <NA>       11 URIAH ST       11 URIAH ST    NEW HAVEN
## 754267              <NA>  181 SOUTH END RD  181 SOUTH END RD    NEW HAVEN
## 754268              <NA>  169 SOUTH END RD  169 SOUTH END RD    NEW HAVEN
## 754270              <NA> 165 SOUTH END RD  165 SOUTH END RD     NEW HAVEN
##        mailing_state assessed_total assessed_land assessed_building
## 754264            CT         552510        552510                0
## 754265            CT         203490         64330           128240
## 754266            CT         199430         61950           137480
## 754267            CT         161000         67200            93800
## 754268            CT         144550         66150            67900
## 754270            CT         131810         61390            70420
##        pre_year_assessed_total appraised_land appraised_building
## 754264                  552510         789300                0
## 754265                  203490          91900           183200
## 754266                  199430          88500           196400
## 754267                  161000          96000           134000
## 754268                  144550          94500            97000
## 754270                  131810          87700           100600
##        appraised_outbuilding appraised_extra_feature valuation_year zone
```

```
## 754264                    0                      NA           2021  RS2
## 754265                11900                      NA           2021  RS2
## 754266                    0                      NA           2021  RS2
## 754267                    0                      NA           2021  RS2
## 754268                15000                      NA           2021  RS2
## 754270                    0                      NA           2021  RS2
##        zone_description model condition condition_description  ayb  eyb
## 754264             <NA>     0      <NA>                  <NA>    0    0
## 754265             <NA>     1         G                  Good 1950 1999
## 754266             <NA>     1         A               Average 1989 2003
## 754267             <NA>     1         G                  Good 1945 1999
## 754268             <NA>     1         A               Average 1940 1989
## 754270             <NA>     1         A               Average 1940 1989
##        living_area effective_area total_rooms number_of_bedroom number_of_baths
## 754264           0              0          NA                NA              NA
## 754265        1475           1706           6                 3               2
## 754266        1792           1996           6                 3               2
## 754267         864           1129           8                 3               1
## 754268        1040           1289           5                 3               1
## 754270        1080           1238           6                 3               1
##        number_of_half_baths occupancy sale_price           sale_date qualified
## 754264                   NA        NA          0 1975-08-05 20:00:00      <NA>
## 754265                    1         1     132000 2017-01-24 19:00:00         U
## 754266                    1         1     140000 2001-12-30 19:00:00         Q
## 754267                    0         1     183000 2004-07-28 20:00:00         Q
## 754268                    0         1          0 2018-10-03 20:00:00         U
## 754270                    1         1          0 2021-02-24 19:00:00         U
##            prior_sale_date prior_book_page prior_sale_price editor edit_date
## 754264                <NA>            <NA>               NA   <NA>      <NA>
## 754265 2016-05-12 20:00:00       9418/0207                0   <NA>      <NA>
## 754266 1997-10-21 20:00:00       5222/0120                0   <NA>      <NA>
## 754267 1992-02-17 19:00:00       4455/0004            90000   <NA>      <NA>
## 754268 1990-11-01 19:00:00       4307/0304                0   <NA>      <NA>
## 754270 2020-05-26 20:00:00      10005/0104                0   <NA>      <NA>
##        collection_year planning_region state_use state_use_description
## 754264            2023   South Central      <NA>          CITY  MDL-00
## 754265            2023   South Central      1010         Single Family
## 754266            2023   South Central      1010         Single Family
## 754267            2023   South Central      1010         Single Family
## 754268            2023   South Central      1010         Single Family
## 754270            2023   South Central      1010         Single Family
##                                    globalid shape_length shape_area
## 754264 {C0B110A2-1E30-4235-BB9D-D68D554110BA}     890.4312 17083.1649
## 754265 {3F400E85-FA62-43AC-AB66-33CE8B941DC1}     190.1392  1797.0473
## 754266 {DB19BE28-C4AB-4BB7-8A22-BAA81389353E}     181.8368  1303.4148
## 754267 {48912D21-5FB7-4F14-9027-09A16FEB90F3}     315.4156  2986.3171
## 754268 {91C7C939-436A-4AE7-B66D-E4669BB0BD0B}     250.3992  2106.8643
## 754270 {526B9FD9-DDBC-4CCE-BCB9-2D46A4B33769}     136.6323   960.4594
```

Number of Bedrooms: While single family may include studios we check whether the variable can be trusted. For instance, we are concerned that NA value might have been translated to 0 values

```r
# checked 4 WARWICK ST #12: it has 1 bath and 1 bed, data has problems
# checked 956 QUINNIPIAC AV: it is barren land
# checked 115 CRANSTON ST: it has 2 bed and 1 bath, data has problems
# checked 532 CHAPEL ST: it has 1 bed, 1 bath, 2 restrooms, data has problems
head(d[d$number_of_bedroom == 0 & !is.na(d$number_of_bedroom), ])
```

```
##        town_name                link                   owner co_owner
## 757585 NEW HAVEN 52070-085 0997 00112             ARNONE JOHN     <NA>
## 757586 NEW HAVEN 52070-085 0997 00113             ARNONE JOHN     <NA>
## 761611 NEW HAVEN 52070-114 1012 00500 RAVENWOOD PROPERTIES LLC     <NA>
## 761786 NEW HAVEN 52070-142 1060 01400        SKOMRO BARBARA M      <NA>
## 762212 NEW HAVEN 52070-207 0543 01100         DASCANIO JOSEPH      <NA>
## 762574 NEW HAVEN 52070-209 0590 00202        TISHS TERRACE LLC     <NA>
##                 location  mailing_address mailing_city mailing_state
## 757585  4 WARWICK ST #12       259 COE AV   EAST HAVEN            CT
## 757586  4 WARWICK ST #13      259 COE AVE   EAST HAVEN            CT
## 761611 956 QUINNIPIAC AV 152 TOWN FARM RD   FARMINGTON            CT
## 761786   115 CRANSTON ST  115 CRANSTON ST    NEW HAVEN            CT
## 762212     532 CHAPEL ST    532 CHAPEL ST    NEW HAVEN            CT
## 762574   667 STATE ST #A     66 HOYT LANE     GUILFORD            CT
##        assessed_total assessed_land assessed_building pre_year_assessed_total
## 757585          19880             0             19880                   19880
## 757586          19880             0             19880                   19880
## 761611           5530             0                 0                    5530
## 761786          67900         38080             29820                   67900
## 762212         245910        104510            133210                  245910
## 762574         274120            70            274050                  274120
##        appraised_land appraised_building appraised_outbuilding
## 757585              0              28400                     0
## 757586              0              28400                     0
## 761611              0                  0                  7900
## 761786          54400              42600                     0
## 762212         149300             190300                 11700
## 762574            100             391500                     0
##        appraised_extra_feature valuation_year zone zone_description model
## 757585                    NA           2021  RM1             <NA>     5
## 757586                    NA           2021  RM1             <NA>     5
## 761611                    NA           2021  RM1             <NA>     5
## 761786                    NA           2021  RS2             <NA>     1
## 762212                    NA           2021  RM2             <NA>     1
## 762574                    NA           2021   BA             <NA>     6
##        condition condition_description  ayb  eyb living_area effective_area
## 757585         G                 Good 1989 2009         420            420
## 757586         G                 Good 1989 2009         420            420
## 761611         A              Average 1987 2005           1              1
## 761786         A              Average 2009 2011         384            384
## 762212         G                 Good 1844 1999        2436           2716
## 762574         G                 Good 1988 2001        3000           3000
##        total_rooms number_of_bedroom number_of_baths number_of_half_baths
## 757585           1                 0             1.0                    0
## 757586           1                 0             1.0                    0
## 761611          NA                 0             0.0                    0
## 761786           0                 0             0.0                    0
```

```
## 762212          NA                0            2.0                   0
## 762574          NA                0            0.5                   0
##        occupancy sale_price           sale_date qualified     prior_sale_date
## 757585         1      35000 2006-07-25 20:00:00         Q 1996-04-02 19:00:00
## 757586         1      34900 2005-06-14 20:00:00         Q 2001-05-24 20:00:00
## 761611         0          0 2011-11-20 19:00:00         U 2007-06-24 20:00:00
## 761786         0          0 1982-10-12 20:00:00         U                <NA>
## 762212         2          0 1986-05-29 20:00:00      <NA>                <NA>
## 762574         1          0 2010-07-27 20:00:00         U 2010-07-27 20:00:00
##        prior_book_page prior_sale_price editor edit_date collection_year
## 757585       4981/0171                0   <NA>      <NA>            2023
## 757586       5861/0241                0   <NA>      <NA>            2023
## 761611       7990/0064                0   <NA>      <NA>            2023
## 761786            <NA>               NA   <NA>      <NA>            2023
## 762212            <NA>               NA   <NA>      <NA>            2023
## 762574       8577/0209           937500   <NA>      <NA>            2023
##        planning_region state_use state_use_description
## 757585   South Central      1020           Condominium
## 757586   South Central      1020           Condominium
## 761611   South Central      1020           Condominium
## 761786   South Central      <NA>    Outbulding  MDL-01
## 762212   South Central      <NA>    OFFICE BLD  MDL-01
## 762574   South Central      3401     OFF CONDO  MDL-06
##                                      globalid shape_length shape_area
## 757585 {78B4132D-A24B-48EC-B3A7-91530E34BD19}     334.7013  3550.0067
## 757586 {FF9B5557-98B1-4771-86ED-DB876F9D1134}     334.7013  3550.0067
## 761611 {A1919646-1C57-41CE-8CFF-59B97A00F1BC}     919.9934 50163.2480
## 761786 {F7A76E2A-0A51-46D4-95BC-EB7831BB6B65}     121.8422   830.1841
## 762212 {1674A82E-9C59-4940-9AAF-1D113286E751}     159.2918  1170.6886
## 762574 {A93E9AFB-8FD2-445A-BBF3-1280E4A859FC}     366.7117  5385.6552
```

```
# Check properties with 0 bathrooms
# Checked 956 QUINNIPIAC AV: is not a house
# Checked 683 STATE ST #I: is a great unit, has 1 bath, data has problems
head(d[d$number_of_baths == 0 & !is.na(d$number_of_baths), ], 10)
```

```
##        town_name                link                    owner
## 761611 NEW HAVEN 52070-114 1012 00500 RAVENWOOD PROPERTIES LLC
## 761786 NEW HAVEN 52070-142 1060 01400        SKOMRO BARBARA M
## 762575 NEW HAVEN 52070-209 0590 00203        TISHS TERRACE LLC
## 762576 NEW HAVEN 52070-209 0590 00204    671 STATE HOLDING LLC
## 762577 NEW HAVEN 52070-209 0590 00205       PALLMAN & COMPANY
## 762578 NEW HAVEN 52070-209 0590 00206             CAITFLO LLC
## 762579 NEW HAVEN 52070-209 0590 00207           677 STATE LLC
## 762580 NEW HAVEN 52070-209 0590 00208     679 STATE STREET LLC
## 762581 NEW HAVEN 52070-209 0590 00209     681 STATE STREET LLC
## 762582 NEW HAVEN 52070-209 0590 00210       BUCKLEY JOHN F JR
##                   co_owner          location     mailing_address mailing_city
## 761611               <NA> 956 QUINNIPIAC AV    152 TOWN FARM RD   FARMINGTON
## 761786               <NA>   115 CRANSTON ST     115 CRANSTON ST    NEW HAVEN
## 762575               <NA>   669 STATE ST #B        66 HOYT LANE     GUILFORD
## 762576               <NA>   671 STATE ST #C         22 ALLEN RD  NORTH HAVEN
## 762577               <NA>   673 STATE ST #D 677 STATE ST UNIT D    NEW HAVEN
## 762578               <NA>   675 STATE ST #E      675 STATE ST E    NEW HAVEN
```

```
## 762579               <NA>   677 STATE ST #F     677 STATE ST #F    NEW HAVEN
## 762580               <NA>   679 STATE ST #G     679 STATE ST #G    NEW HAVEN
## 762581               <NA>   681 STATE ST #H       681  STATE ST    NEW HAVEN
## 762582 C/O BUCKLEY & WYNNE   683 STATE ST #I        685 STATE ST    NEW HAVEN
##        mailing_state assessed_total assessed_land assessed_building
## 761611            CT           5530             0                 0
## 761786            CT          67900         38080             29820
## 762575            CT         418670            70            418600
## 762576            CT         337820            70            337750
## 762577            CT         176120            70            176050
## 762578            CT         176120            70            176050
## 762579            CT         176120            70            176050
## 762580            CT         256970            70            256900
## 762581            CT         256970            70            256900
## 762582            CT         256970            70            256900
##        pre_year_assessed_total appraised_land appraised_building
## 761611                    5530             0                 0
## 761786                   67900          54400             42600
## 762575                  418670            100            598000
## 762576                  337820            100            482500
## 762577                  176120            100            251500
## 762578                  176120            100            251500
## 762579                  176120            100            251500
## 762580                  256970            100            367000
## 762581                  256970            100            367000
## 762582                  256970            100            367000
##        appraised_outbuilding appraised_extra_feature valuation_year zone
## 761611                  7900                      NA           2021  RM1
## 761786                     0                      NA           2021  RS2
## 762575                     0                      NA           2021   BA
## 762576                     0                      NA           2021   BA
## 762577                     0                      NA           2021   BA
## 762578                     0                      NA           2021   BA
## 762579                     0                      NA           2021   BA
## 762580                     0                      NA           2021   BA
## 762581                     0                      NA           2021   BA
## 762582                     0                      NA           2021   BA
##        zone_description model condition condition_description  ayb  eyb
## 761611             <NA>     5         A               Average 1987 2005
## 761786             <NA>     1         A               Average 2009 2011
## 762575             <NA>     6         A               Average 1988 1996
## 762576             <NA>     6         A               Average 1988 1996
## 762577             <NA>     6         A               Average 1988 1996
## 762578             <NA>     6         A               Average 1988 1996
## 762579             <NA>     6         A               Average 1988 1996
## 762580             <NA>     6         A               Average 1988 1996
## 762581             <NA>     6         A               Average 1988 1996
## 762582             <NA>     6         A               Average 1988 1996
##        living_area effective_area total_rooms number_of_bedroom number_of_baths
## 761611           1              1          NA                 0               0
## 761786         384            384           0                 0               0
## 762575        5000           5000          NA                 0               0
## 762576        4000           4000          NA                 0               0
## 762577        2000           2000          NA                 0               0
```

```
## 762578        2000           2000          NA                0              0
## 762579        2000           2000          NA                0              0
## 762580        3000           3000          NA                0              0
## 762581        3000           3000          NA                0              0
## 762582        3000           3000          NA                0              0
##        number_of_half_baths occupancy sale_price           sale_date qualified
## 761611                    0         0          0 2011-11-20 19:00:00         U
## 761786                    0         0          0 1982-10-12 20:00:00         U
## 762575                    1         1          0 2010-07-27 20:00:00         U
## 762576                    0         1          0 2007-03-14 20:00:00         U
## 762577                    0         1          0 1990-06-07 20:00:00      <NA>
## 762578                    0         1     130000 2001-03-29 19:00:00         U
## 762579                    0         1     279000 2022-04-13 20:00:00         Q
## 762580                    1         1     297500 2001-07-17 20:00:00         Q
## 762581                    0         1          0 2011-09-20 20:00:00         U
## 762582                    0         1     450000 1988-09-01 20:00:00         Q
##            prior_sale_date prior_book_page prior_sale_price editor edit_date
## 761611 2007-06-24 20:00:00       7990/0064                0   <NA>      <NA>
## 761786                <NA>            <NA>               NA   <NA>      <NA>
## 762575 2010-07-27 20:00:00       8577/0209           937500   <NA>      <NA>
## 762576 1996-09-10 20:00:00       5042/0234                0   <NA>      <NA>
## 762577                <NA>            <NA>               NA   <NA>      <NA>
## 762578 1990-04-11 20:00:00       4235/0114                0   <NA>      <NA>
## 762579 2000-01-11 19:00:00       5612/0004                0   <NA>      <NA>
## 762580 1988-05-01 20:00:00       3882/0294           450000   <NA>      <NA>
## 762581 1995-04-05 20:00:00       4852/0113                0   <NA>      <NA>
## 762582                <NA>            <NA>               NA   <NA>      <NA>
##        collection_year planning_region state_use state_use_description
## 761611            2023   South Central      1020           Condominium
## 761786            2023   South Central      <NA>    Outbulding  MDL-01
## 762575            2023   South Central      3401     OFF CONDO  MDL-06
## 762576            2023   South Central      3401     OFF CONDO  MDL-06
## 762577            2023   South Central      3401     OFF CONDO  MDL-06
## 762578            2023   South Central      3401     OFF CONDO  MDL-06
## 762579            2023   South Central      3401     OFF CONDO  MDL-06
## 762580            2023   South Central      3401     OFF CONDO  MDL-06
## 762581            2023   South Central      3401     OFF CONDO  MDL-06
## 762582            2023   South Central      3401     OFF CONDO  MDL-06
##                                    globalid shape_length shape_area
## 761611 {A1919646-1C57-41CE-8CFF-59B97A00F1BC}     919.9934 50163.2480
## 761786 {F7A76E2A-0A51-46D4-95BC-EB7831BB6B65}     121.8422   830.1841
## 762575 {1CE56061-7AFD-4E58-A6F3-A2EBCE67E255}     366.7117  5385.6552
## 762576 {A66C021A-47D1-4433-81F3-9082DCA34616}     366.7117  5385.6552
## 762577 {E8BFFE4F-B7D0-45A5-9850-B606BB6B0B55}     366.7117  5385.6552
## 762578 {BDB62FD9-1139-4A7F-AC62-7940E17C7580}     366.7117  5385.6552
## 762579 {8E526E30-C099-4517-96F7-C64FB00F00F8}     366.7117  5385.6552
## 762580 {D9DA0082-51DB-4282-BB4B-9B27B87B2FB1}     366.7117  5385.6552
## 762581 {A21CDDB8-04DC-46CB-B37E-26F40D2EC7B0}     366.7117  5385.6552
## 762582 {5B991AFF-6026-4D8E-9D46-37E8967BF85D}     366.7117  5385.6552
```

```r
# Checked 200 ORCHARD ST #408: is a great unit, has 1 bath, data has problems
# Checked 1395 CHAPEL ST: it's a dental building, it should have a bath!
tail(d[d$number_of_baths == 0 & !is.na(d$number_of_baths), ], 10)
```

```
##        town_name                link                       owner
## 774448 NEW HAVEN 52070-315 1288 02032 YALE-NEW HAVEN HOSPITAL INC
## 774449 NEW HAVEN 52070-315 1288 02033 YALE-NEW HAVEN HOSPITAL INC
## 774450 NEW HAVEN 52070-315 1288 02034 YALE-NEW HAVEN HOSPITAL INC
## 774451 NEW HAVEN 52070-315 1288 02035 YALE-NEW HAVEN HOSPITAL INC
## 774452 NEW HAVEN 52070-315 1288 02036 YALE-NEW HAVEN HOSPITAL INC
## 774921 NEW HAVEN 52070-316 0243 02600         ALBANIA DENTAL LLC
## 777198 NEW HAVEN 52070-310 0091 02000          S&E APARTMENTS LLC
## 779853 NEW HAVEN 52070-407 1241 00300     LISA BETH REALTY CO INC
## 780965 NEW HAVEN 52070-033 0868 01500           HELLYAR E MICHAEL
## 781175 NEW HAVEN 52070-033 0868 00900        NEW HAVEN YACHT CLUB
##                    co_owner            location mailing_address mailing_city
## 774448                <NA> 200 ORCHARD ST #404      20 YORK ST    NEW HAVEN
## 774449                <NA> 200 ORCHARD ST #405      20 YORK ST    NEW HAVEN
## 774450                <NA> 200 ORCHARD ST #406      20 YORK ST    NEW HAVEN
## 774451                <NA> 200 ORCHARD ST #407      20 YORK ST    NEW HAVEN
## 774452                <NA> 200 ORCHARD ST #408      20 YORK ST    NEW HAVEN
## 774921                <NA>      1395 CHAPEL ST  1395 CHAPEL ST    NEW HAVEN
## 777198                <NA>        26 DOWNES ST   228 COLONY RD    NEW HAVEN
## 779853 C/O ROHINSKY IRVING     49 EDGEWOOD #WY  49 EDGEWOOD WY    NEW HAVEN
## 780965                <NA>         132 COVE ST    96 FORBES PL   EAST HAVEN
## 781175                <NA>         156 COVE ST   P.O. BOX 8163    NEW HAVEN
##        mailing_state assessed_total assessed_land assessed_building
## 774448            CT          97650             0             97650
## 774449            CT          99680             0             99680
## 774450            CT          95060             0             95060
## 774451            CT          65100             0             65100
## 774452            CT         100520             0            100520
## 774921            CT         169120         68600             94360
## 777198            CT         119420         35490             83930
## 779853            CT         496440         98070            398370
## 780965            CT         238280        205450             32830
## 781175            CT         324800        241360             78400
##        pre_year_assessed_total appraised_land appraised_building
## 774448                   97650              0             139500
## 774449                   99680              0             142400
## 774450                   95060              0             135800
## 774451                   65100              0              93000
## 774452                  100520              0             143600
## 774921                  169120          98000             134800
## 777198                  119420          50700             119900
## 779853                  496440         140100             569100
## 780965                  238280         293500              46900
## 781175                  324800         344800             112000
##        appraised_outbuilding appraised_extra_feature valuation_year zone
## 774448                     0                      NA           2021  RM2
## 774449                     0                      NA           2021  RM2
## 774450                     0                      NA           2021  RM2
## 774451                     0                      NA           2021  RM2
## 774452                     0                      NA           2021  RM2
## 774921                  8800                      NA           2021   RO
## 777198                     0                      NA           2021  RM2
## 779853                     0                      NA           2021  RS2
## 780965                     0                      NA           2021  RS2
```

```
## 781175                 7200                      NA           2021  RS2
##        zone_description model condition condition_description  ayb  eyb
## 774448             <NA>     6         G                  Good 1906 1986
## 774449             <NA>     6         G                  Good 1906 1986
## 774450             <NA>     6         G                  Good 1906 1986
## 774451             <NA>     6         G                  Good 1906 1986
## 774452             <NA>     6         G                  Good 1906 1986
## 774921             <NA>     1         G                  Good 1900 1999
## 777198             <NA>     1         F                     F 1910 1976
## 779853             <NA>     3         F                     F 1900 1976
## 780965             <NA>     1         A               Average 1920 1989
## 781175             <NA>     1         A               Average 1910 1989
##        living_area effective_area total_rooms number_of_bedroom number_of_baths
## 774448        1514           1514          NA                 0               0
## 774449        1549           1549          NA                 0               0
## 774450        1470           1470          NA                 0               0
## 774451         954            954          NA                 0               0
## 774452        1564           1564          NA                 0               0
## 774921        2061           2257          NA                 2               0
## 777198        1848           2068           8                 4               0
## 779853       10962          11192          NA                 0               0
## 780965         540            609           2                 0               0
## 781175         830            975           2                 0               0
##        number_of_half_baths occupancy sale_price           sale_date qualified
## 774448                    0         1          0 2012-09-10 20:00:00         U
## 774449                    0         1          0 2012-09-10 20:00:00         U
## 774450                    0         1          0 2012-09-10 20:00:00         U
## 774451                    0         1          0 2012-09-10 20:00:00         U
## 774452                    0         1          0 2012-09-10 20:00:00         U
## 774921                    0         2      92000 2000-08-15 20:00:00         Q
## 777198                    2         1          0 2018-06-07 20:00:00         U
## 779853                    0         9          0 1963-02-26 19:00:00         U
## 780965                    1         1     225000 2015-01-14 19:00:00         Q
## 781175                    2         1          0 2003-02-20 19:00:00         U
##            prior_sale_date prior_book_page prior_sale_price editor edit_date
## 774448 1999-09-28 20:00:00       5565/0018                0   <NA>      <NA>
## 774449 1999-09-28 20:00:00       5565/0018                0   <NA>      <NA>
## 774450 1999-09-28 20:00:00       5565/0018                0   <NA>      <NA>
## 774451 1999-09-28 20:00:00       5565/0018                0   <NA>      <NA>
## 774452 1999-09-28 20:00:00       5565/0018                0   <NA>      <NA>
## 774921 1999-12-29 19:00:00       5606/0314                0   <NA>      <NA>
## 777198 2017-05-11 20:00:00       9569/0338                0   <NA>      <NA>
## 779853                <NA>            <NA>               NA   <NA>      <NA>
## 780965 1996-03-28 19:00:00       4979/0014                0   <NA>      <NA>
## 781175 2003-02-20 19:00:00       6341/0293                0   <NA>      <NA>
##        collection_year planning_region state_use state_use_description
## 774448            2023   South Central      9059    EX MED CND  MDL-06
## 774449            2023   South Central      9059    EX MED CND  MDL-06
## 774450            2023   South Central      9059    EX MED CND  MDL-06
## 774451            2023   South Central      9059    EX MED CND  MDL-06
## 774452            2023   South Central      9059    EX MED CND  MDL-06
## 774921            2023   South Central      <NA>     MIXED USE  MDL-01
## 777198            2023   South Central      1010         Single Family
## 779853            2023   South Central      1111   APT5 - 12 R  MDL-03
```

```
## 780965            2023   South Central      1013             SFR Water
## 781175            2023   South Central      1013             SFR Water
##                                      globalid shape_length shape_area
## 774448 {D9B01FA4-53F1-410A-BF66-65E2D951BAD6}     462.4735  9894.6960
## 774449 {C7026A6C-877B-4BA6-9E09-E4C13334D661}     462.4735  9894.6960
## 774450 {B79827F7-B9B2-4E59-B371-FA2FBE22BE25}     462.4735  9894.6960
## 774451 {149A157C-AB0E-4245-99CF-8B00C4207D05}     462.4735  9894.6960
## 774452 {75719626-B2D9-42CC-9B5B-0355B57C2880}     462.4735  9894.6960
## 774921 {2FF431AF-5202-4723-A901-2A438F3BA407}     159.3500  1164.4685
## 777198 {127388B3-879C-48A5-B57A-DE1F6447EE56}     109.4460   641.1594
## 779853 {2707BA8B-92DC-42DB-9AED-44FD1E46B700}     272.4178  4467.6087
## 780965 {992D5394-F60D-4E07-BCFF-4159A199F382}     218.6909  1562.8521
## 781175 {A81D1ADF-A8C8-49E0-B4FA-E475838FE2C2}     291.0912  4917.7748
```

```
sum(d$number_of_baths == 0, na.rm = T)
```

```
## [1] 189
```

```
# Remove entries with 0 bedrooms and bathrooms as they don't meet criteria
sum(is.na(d$number_of_bedroom == 0))
```

```
## [1] 6047
```

```
d <- d[d$number_of_bedroom != 0 | is.na(d$number_of_bedroom), ]
sum(is.na(d$number_of_baths == 0))
```

```
## [1] 6045
```

```
d <- d[d$number_of_baths != 0 | is.na(d$number_of_baths), ]
nrow(d) # Check on the number of properties
```

```
## [1] 20959
```

### Assessed Total: Price Restrictions

Set price limits for properties: below $50,000 likely indicate poor conditions, while those above $10 million may not reflect typical single-family homes. This step refines the dataset by removing extreme values.

### Check properties priced below $50,000

```
# Check properties priced below $50,000
sum(d$assessed_total < 50000, na.rm = T)
```

```
## [1] 2493
```

```
# checked 1227 DEAN ST: looks like a nowhere
head(d[d$assessed_total < 50000 & !is.na(d$assessed_total), ])
```

```
##        town_name                 link                          owner
## 754278 NEW HAVEN 52070-014 0853 01200               CITY OF NEW HAVEN
## 754321 NEW HAVEN 52070-016 1303 00100            ABBAGNARO EST JOSEPH
## 754322 NEW HAVEN 52070-016 1303 00200               TOBIA CHRISTOPHER
## 754323 NEW HAVEN 52070-016 1303 00300       RYAN MARIA, LUZZI CHERYL &
## 754324 NEW HAVEN 52070-016 1303 00400        ADINOLFI SALVATORE M JR &
## 754325 NEW HAVEN 52070-016 1303 00500 ADINOLFI SALVATORE & BETTE JAN
##                      co_owner      location   mailing_address mailing_city
## 754278                   <NA>  SOUTH END RD     165 CHURCH ST    NEW HAVEN
## 754321                   <NA>  1209 DEAN ST 265 LIGHTHOUSE RD    NEW HAVEN
## 754322                   <NA>  1227 DEAN ST      1224 DEAN ST    NEW HAVEN
## 754323        RYAN THOMAS R JR  1237 DEAN ST      1236 DEAN ST    NEW HAVEN
## 754324 FOSS-ADINOLFI LAURA JEAN 1251 DEAN ST       1244 DEAN ST    NEW HAVEN
## 754325                   <NA>       DEAN ST      1254 DEAN ST    NEW HAVEN
##        mailing_state assessed_total assessed_land assessed_building
## 754278            CT           2940          2940                 0
## 754321            CT           5880          5880                 0
## 754322            CT           2940          2940                 0
## 754323            CT           2940          2940                 0
## 754324            CT           2940          2940                 0
## 754325            CT           2940          2940                 0
##        pre_year_assessed_total appraised_land appraised_building
## 754278                    2940           4200                  0
## 754321                    5880           8400                  0
## 754322                    2940           4200                  0
## 754323                    2940           4200                  0
## 754324                    2940           4200                  0
## 754325                    2940           4200                  0
##        appraised_outbuilding appraised_extra_feature valuation_year zone
## 754278                     0                      NA           2021  RS2
## 754321                     0                      NA           2021  RS2
## 754322                     0                      NA           2021  RS2
## 754323                     0                      NA           2021  RS2
## 754324                     0                      NA           2021  RS2
## 754325                     0                      NA           2021  RS2
##        zone_description model condition condition_description ayb eyb
## 754278             <NA>     0      <NA>                  <NA>   0   0
## 754321             <NA>     0      <NA>                  <NA>   0   0
## 754322             <NA>     0      <NA>                  <NA>   0   0
## 754323             <NA>     0      <NA>                  <NA>   0   0
## 754324             <NA>     0      <NA>                  <NA>   0   0
## 754325             <NA>     0      <NA>                  <NA>   0   0
##        living_area effective_area total_rooms number_of_bedroom number_of_baths
## 754278           0              0          NA                NA              NA
## 754321           0              0          NA                NA              NA
## 754322           0              0          NA                NA              NA
## 754323           0              0          NA                NA              NA
## 754324           0              0          NA                NA              NA
## 754325           0              0          NA                NA              NA
##        number_of_half_baths occupancy sale_price           sale_date qualified
## 754278                   NA        NA          0 1968-04-22 19:00:00      <NA>
## 754321                   NA        NA          0 1980-12-16 19:00:00      <NA>
## 754322                   NA        NA          0 2016-06-29 20:00:00         U
## 754323                   NA        NA          0 2017-03-16 20:00:00         U
```

```
## 754324                   NA        NA          0 2019-11-18 19:00:00         U
## 754325                   NA        NA          0 1990-05-01 20:00:00      <NA>
##           prior_sale_date prior_book_page prior_sale_price editor edit_date
## 754278               <NA>            <NA>               NA   <NA>      <NA>
## 754321               <NA>            <NA>               NA   <NA>      <NA>
## 754322 2009-09-29 20:00:00       8446/0001           149350   <NA>      <NA>
## 754323 2015-12-28 19:00:00       9366/0276                0   <NA>      <NA>
## 754324 2019-03-11 20:00:00       9825/0035           170000   <NA>      <NA>
## 754325               <NA>            <NA>               NA   <NA>      <NA>
##        collection_year planning_region state_use state_use_description
## 754278            2023   South Central      <NA>     MUNICIPAL  MDL-00
## 754321            2023   South Central      1320            VAC UN BLD
## 754322            2023   South Central      1320            VAC UN BLD
## 754323            2023   South Central      1320            VAC UN BLD
## 754324            2023   South Central      1320            VAC UN BLD
## 754325            2023   South Central      1320            VAC UN BLD
##                                    globalid shape_length shape_area
## 754278 {9C353200-75B4-40A8-B267-DAA1B73F4228}    128.31465   318.2036
## 754321 {4CB4B2A6-D36E-4AB0-AE64-BC3CFCAB31EB}    112.95729   564.0061
## 754322 {15FAAD69-EBD8-4316-AA58-E8F272DBD70E}     68.23209   272.8196
## 754323 {0FB33CE9-8FCE-418B-90CA-43F7367F8EB8}     66.47178   261.2075
## 754324 {FA671F41-62FA-4CAB-A3B3-810C61425580}     67.05733   264.5538
## 754325 {F4D4A2B1-FCA3-420A-9454-5EA41A1C1978}     67.82665   274.1638
```

```
# checked 747 WASHINGTON AV: it's a very old deli store
tail(d[d$assessed_total < 50000 & !is.na(d$assessed_total), ])
```

```
##        town_name                 link                                    owner
## 781141 NEW HAVEN 52070-034 0852 01400                NEW HAVEN LAND TRUST INC
## 781179 NEW HAVEN 52070-030 0871 00100                       CITY OF NEW HAVEN
## 781282 NEW HAVEN 52070-304 0040 00700                       CITY OF NEW HAVEN
## 781283 NEW HAVEN 52070-304 0040 00800 CITY OF NEW HAVEN FOR BOARD OF EDUCATION
## 781285 NEW HAVEN 52070-304 0040 00901                    STATE OF CONNECTICUT
## 781288 NEW HAVEN 52070-305 0068 01901                    STATE OF CONNECTICUT
##                          co_owner              location    mailing_address
## 781141                       <NA> LIGHTHOUSE POINT TER  817 GRAND AVE #102
## 781179                       <NA>        54 TOWNSEND AV     165 CHRUCH ST.
## 781282                       <NA>     745 WASHINGTON AV      165 CHURCH ST
## 781283                       <NA>     747 WASHINGTON AV      165 CHURCH ST
## 781285                       <NA>    ELLA T GRASSO BLVD      P.O. DRAWER A
## 781288 C/O DEPT OF TRANSPORTATION    ELLA T GRASSO BLVD      PO BOX 317546
##        mailing_city mailing_state assessed_total assessed_land
## 781141    NEW HAVEN            CT           6580          6580
## 781179    NEW HAVEN            CT           6370          6370
## 781282    NEW HAVEN            CT          23100         23100
## 781283    NEW HAVEN            CT          23100         23100
## 781285 WETHERSFIELD            CT           2380          2380
## 781288    NEWINGTON            CT          44380         44380
##        assessed_building pre_year_assessed_total appraised_land
## 781141                 0                    6580           9400
## 781179                 0                    6370           9100
## 781282                 0                   23100          33000
## 781283                 0                   23100          33000
## 781285                 0                    2380           3400
```

```
## 781288                 0                   44380          63400
##        appraised_building appraised_outbuilding appraised_extra_feature
## 781141                  0                     0                      NA
## 781179                  0                     0                      NA
## 781282                  0                     0                      NA
## 781283                  0                     0                      NA
## 781285                  0                     0                      NA
## 781288                  0                     0                      NA
##        valuation_year zone zone_description model condition
## 781141           2021  RS2             <NA>     0      <NA>
## 781179           2021 PARK             <NA>     0      <NA>
## 781282           2021  RM2             <NA>     0      <NA>
## 781283           2021  RM2             <NA>     0      <NA>
## 781285           2021   IL             <NA>     0      <NA>
## 781288           2021   IL             <NA>     0      <NA>
##        condition_description ayb eyb living_area effective_area total_rooms
## 781141                  <NA>   0   0           0              0          NA
## 781179                  <NA>   0   0           0              0          NA
## 781282                  <NA>   0   0           0              0          NA
## 781283                  <NA>   0   0           0              0          NA
## 781285                  <NA>   0   0           0              0          NA
## 781288                  <NA>   0   0           0              0          NA
##        number_of_bedroom number_of_baths number_of_half_baths occupancy
## 781141                NA              NA                   NA        NA
## 781179                NA              NA                   NA        NA
## 781282                NA              NA                   NA        NA
## 781283                NA              NA                   NA        NA
## 781285                NA              NA                   NA        NA
## 781288                NA              NA                   NA        NA
##        sale_price           sale_date qualified     prior_sale_date
## 781141          0 2003-08-13 20:00:00         U 2003-05-07 20:00:00
## 781179          0 2004-05-23 20:00:00         U                <NA>
## 781282          0 2002-11-18 19:00:00         U 1998-07-01 20:00:00
## 781283          0 2003-02-12 19:00:00         U 2001-07-24 20:00:00
## 781285          0 2000-07-26 20:00:00         U                <NA>
## 781288          0 2000-08-10 20:00:00         U                <NA>
##        prior_book_page prior_sale_price editor edit_date collection_year
## 781141       6416/0235                0   <NA>      <NA>            2023
## 781179            0/0                0   <NA>      <NA>            2023
## 781282       5336/0012                0   <NA>      <NA>            2023
## 781283       5904/0242            39000   <NA>      <NA>            2023
## 781285            <NA>               NA   <NA>      <NA>            2023
## 781288            <NA>               NA   <NA>      <NA>            2023
##        planning_region state_use state_use_description
## 781141   South Central      <NA>    NON-PROFIT  MDL-00
## 781179   South Central      <NA>     MUNICIPAL  MDL-00
## 781282   South Central      <NA>     MUNICIPAL  MDL-00
## 781283   South Central      <NA>     MUNICIPAL  MDL-00
## 781285   South Central      <NA>     STATE ADM  MDL-00
## 781288   South Central      <NA>     STATE ADM  MDL-00
##                                    globalid shape_length shape_area
## 781141 {53DDB3AB-289A-4C23-9F81-A5D5486C5781}    171.04414 1451.61712
## 781179 {6B30E38B-8E69-451B-9287-ED47073F7781}     71.89174  136.96965
## 781282 {E79AFBD8-725A-4144-B7BE-ACD484266639}    113.45396  629.09818
```

```
## 781283 {75674FF9-E678-498E-856B-C5A5F1211B88}    115.29363  670.02975
## 781285 {47CDFD19-6B36-47F4-A0BE-09A9AF0A1948}     36.74541   63.64341
## 781288 {AB0D597A-E81D-4FC9-A112-377DCBF45FF9}    196.75735  814.36019
```

```
# Filter properties priced below $50,000
d <- d[d$assessed_total > 50000, ]
head(d)
```

```
##        town_name                 link                     owner
## 754264 NEW HAVEN 52070-013 0853 00500 CITY OF NEW HAVEN AIRPORT
## 754265 NEW HAVEN 52070-014 0853 00100  PEREZ-RAMIREZ NOE MARTIN
## 754266 NEW HAVEN 52070-014 0853 00101 RODRIGUEZ WILLIAM & LYSIE
## 754267 NEW HAVEN 52070-014 0853 00200             GUEST CRAIG C
## 754268 NEW HAVEN 52070-014 0853 00300  MEADOWS BERNADETTE (EST)
## 754270 NEW HAVEN 52070-014 0853 00400        KILFEATHER FRANCIS
##                  co_owner          location   mailing_address mailing_city
## 754264 CITY OF NEW HAVEN   75 SOUTH END RD     165 CHURCH ST    NEW HAVEN
## 754265              <NA>  199 SOUTH END RD  199 SOUTH END RD    NEW HAVEN
## 754266              <NA>       11 URIAH ST       11 URIAH ST    NEW HAVEN
## 754267              <NA>  181 SOUTH END RD  181 SOUTH END RD    NEW HAVEN
## 754268              <NA>  169 SOUTH END RD  169 SOUTH END RD    NEW HAVEN
## 754270              <NA> 165 SOUTH END RD  165 SOUTH END RD     NEW HAVEN
##        mailing_state assessed_total assessed_land assessed_building
## 754264            CT         552510        552510                0
## 754265            CT         203490         64330           128240
## 754266            CT         199430         61950           137480
## 754267            CT         161000         67200            93800
## 754268            CT         144550         66150            67900
## 754270            CT         131810         61390            70420
##        pre_year_assessed_total appraised_land appraised_building
## 754264                  552510         789300                 0
## 754265                  203490          91900            183200
## 754266                  199430          88500            196400
## 754267                  161000          96000            134000
## 754268                  144550          94500             97000
## 754270                  131810          87700            100600
##        appraised_outbuilding appraised_extra_feature valuation_year zone
## 754264                     0                     NA           2021  RS2
## 754265                 11900                     NA           2021  RS2
## 754266                     0                     NA           2021  RS2
## 754267                     0                     NA           2021  RS2
## 754268                 15000                     NA           2021  RS2
## 754270                     0                     NA           2021  RS2
##        zone_description model condition condition_description  ayb  eyb
## 754264             <NA>     0      <NA>                 <NA>    0    0
## 754265             <NA>     1         G                 Good 1950 1999
## 754266             <NA>     1         A              Average 1989 2003
## 754267             <NA>     1         G                 Good 1945 1999
## 754268             <NA>     1         A              Average 1940 1989
## 754270             <NA>     1         A              Average 1940 1989
##        living_area effective_area total_rooms number_of_bedroom number_of_baths
## 754264           0              0          NA                NA              NA
## 754265        1475           1706           6                 3               2
## 754266        1792           1996           6                 3               2
```

```
## 754267         864           1129           8                3               1
## 754268        1040           1289           5                3               1
## 754270        1080           1238           6                3               1
##        number_of_half_baths occupancy sale_price           sale_date qualified
## 754264                   NA        NA          0 1975-08-05 20:00:00      <NA>
## 754265                    1         1     132000 2017-01-24 19:00:00         U
## 754266                    1         1     140000 2001-12-30 19:00:00         Q
## 754267                    0         1     183000 2004-07-28 20:00:00         Q
## 754268                    0         1          0 2018-10-03 20:00:00         U
## 754270                    1         1          0 2021-02-24 19:00:00         U
##            prior_sale_date prior_book_page prior_sale_price editor edit_date
## 754264                <NA>           <NA>               NA   <NA>      <NA>
## 754265 2016-05-12 20:00:00       9418/0207                0   <NA>      <NA>
## 754266 1997-10-21 20:00:00       5222/0120                0   <NA>      <NA>
## 754267 1992-02-17 19:00:00       4455/0004            90000   <NA>      <NA>
## 754268 1990-11-01 19:00:00       4307/0304                0   <NA>      <NA>
## 754270 2020-05-26 20:00:00      10005/0104                0   <NA>      <NA>
##        collection_year planning_region state_use state_use_description
## 754264            2023   South Central      <NA>          CITY  MDL-00
## 754265            2023   South Central      1010         Single Family
## 754266            2023   South Central      1010         Single Family
## 754267            2023   South Central      1010         Single Family
## 754268            2023   South Central      1010         Single Family
## 754270            2023   South Central      1010         Single Family
##                                    globalid shape_length shape_area
## 754264 {C0B110A2-1E30-4235-BB9D-D68D554110BA}     890.4312 17083.1649
## 754265 {3F400E85-FA62-43AC-AB66-33CE8B941DC1}     190.1392  1797.0473
## 754266 {DB19BE28-C4AB-4BB7-8A22-BAA81389353E}     181.8368  1303.4148
## 754267 {48912D21-5FB7-4F14-9027-09A16FEB90F3}     315.4156  2986.3171
## 754268 {91C7C939-436A-4AE7-B66D-E4669BB0BD0B}     250.3992  2106.8643
## 754270 {526B9FD9-DDBC-4CCE-BCB9-2D46A4B33769}     136.6323   960.4594
```

### Check properties priced over $1M

```
# Check properties priced over $1M
sum(d$assessed_total > 10000000, na.rm = T)
```

```
## [1] 217
```

```
head(d[d$assessed_total > 10000000 & !is.na(d$assessed_total),])
```

```
##        town_name                 link                         owner
## 755074 NEW HAVEN 52070-028 0900 00100     CITY OF NEW HAVEN AIRPORT
## 755853 NEW HAVEN 52070-044 0902 01300      CITY OF NEW HAVEN SCHOOL
## 755971 NEW HAVEN 52070-052 0950 00200    PSEG POWER CONNECTICUT LLC
## 755972 NEW HAVEN 52070-052 0950 00400      CITY OF NEW HAVEN (WPCA)
## 756700 NEW HAVEN 52070-063 0950 00500        TRITON TERMINALING LLC
## 757789 NEW HAVEN 52070-090 0998 00100 CITY OF NEW HAVEN FOR THE BOE
##                 co_owner             location    mailing_address mailing_city
## 755074 CITY OF NEW HAVEN         155 BURR ST       165 CHURCH ST    NEW HAVEN
## 755853              <NA>      480 TOWNSEND AV      165 CHURCH ST    NEW HAVEN
## 755971              <NA>   600 CONNECTICUT AV 80 PARK PLAZA T-6B       NEWARK
## 755972              <NA>  345 EAST SHORE PKWY      165 CHURCH ST    NEW HAVEN
```

```
## 756700              <NA> 481 EAST SHORE PKWY    910 LOUISIANA ST      HOUSTON
## 757789              <NA>      15 LEXINGTON AV    15 LEXINGTON AV    NEW HAVEN
##        mailing_state assessed_total assessed_land assessed_building
## 755074            CT       37738540      34973120            739060
## 755853            CT       14309120        133350          13934130
## 755971            NJ       41752270      11257610          26227880
## 755972            CT      203993230      18835250           4590600
## 756700            TX       10926580       3902640           1040130
## 757789            CT       35554890        668360          34645170
##        pre_year_assessed_total appraised_land appraised_building
## 755074                37402680       49961600            1055800
## 755853                14309120         190500           19905900
## 755971                41752270       16082300           37468400
## 755972               203993230       26907500            6558000
## 756700                10926580        5575200            1485900
## 757789                35554890         954800           49493100
##        appraised_outbuilding appraised_extra_feature valuation_year    zone
## 755074               2880300                      NA           2021 AIRPORT
## 755853                 83700                      NA           2021     RS2
## 755971               5892200                      NA           2021      IH
## 755972             257953400                      NA           2021 IH/PARK
## 756700               8516500                      NA           2021  IH/RM1
## 757789                 92500                      NA           2021     RM1
##        zone_description model condition condition_description  ayb  eyb
## 755074             <NA>    94        VG             Very Good 1931 1994
## 755853             <NA>    94         G                  Good 1924 1986
## 755971             <NA>    96         E             Excellent 1975 2004
## 755972             <NA>    96         A               Average 1992 2001
## 756700             <NA>    96         A               Average 1953 1981
## 757789             <NA>    94         G                  Good 2007 2016
##        living_area effective_area total_rooms number_of_bedroom number_of_baths
## 755074        8655           9414          NA                NA              NA
## 755853       96449          96449          NA                NA              NA
## 755971      117066         119875          NA                NA              NA
## 755972       30234          35334          NA                NA              NA
## 756700        5596           7396          NA                NA              NA
## 757789       85679          85679          NA                NA              NA
##        number_of_half_baths occupancy sale_price           sale_date qualified
## 755074                   NA         1          0 1899-12-31 19:00:00         U
## 755853                   NA         1          0                <NA>      <NA>
## 755971                   NA         1          0 2003-02-20 19:00:00         U
## 755972                   NA         1          0                <NA>      <NA>
## 756700                   NA         1    6919750 2017-05-24 20:00:00         U
## 757789                   NA         1          0 2005-04-28 20:00:00         U
##            prior_sale_date prior_book_page prior_sale_price editor edit_date
## 755074                <NA>            <NA>               NA   <NA>      <NA>
## 755853                <NA>            <NA>               NA   <NA>      <NA>
## 755971 1999-04-18 20:00:00       5483/0007                0   <NA>      <NA>
## 755972                <NA>            <NA>               NA   <NA>      <NA>
## 756700 2000-05-11 20:00:00       5668/0143          2100000   <NA>      <NA>
## 757789 1966-05-02 20:00:00             0/0                0   <NA>      <NA>
##        collection_year planning_region state_use state_use_description
## 755074            2023   South Central      9020          CITY  MDL-94
## 755853            2023   South Central      9033    MUN SCHOOL  MDL-94
```

```
## 755971            2023   South Central      4220            ELEC PLANT
## 755972            2023   South Central      <NA>     MUNICIPAL  MDL-96
## 756700            2023   South Central      4210     TANKS LNG  MDL-96
## 757789            2023   South Central      9026    CITY SCHOO  MDL-94
##                                      globalid shape_length shape_area
## 755074 {FC3F3972-5147-449D-A050-3C9CD387C862}    4169.8397  722662.86
## 755853 {4E7BC857-B8A6-481D-A3DF-8AF5E5783B57}     438.3972   11451.07
## 755971 {4E264C42-D2AB-4194-877E-A4598700B721}    2925.2974  504006.86
## 755972 {5256C08B-332C-4603-B8B8-CCBCA651FFD7}    1374.1364   75101.00
## 756700 {85890434-4336-4566-B855-0D8C37A6A0F8}    2811.4075  271320.92
## 757789 {A0BE3CF4-FA3D-4FCA-AFF5-00C085F93047}    1475.9604  105125.72
```

```
# Filter properties priced over $1M
d <- d[d$assessed_total < 10000000, ]
head(d)
```

```
##        town_name                 link                     owner
## 754264 NEW HAVEN 52070-013 0853 00500 CITY OF NEW HAVEN AIRPORT
## 754265 NEW HAVEN 52070-014 0853 00100  PEREZ-RAMIREZ NOE MARTIN
## 754266 NEW HAVEN 52070-014 0853 00101 RODRIGUEZ WILLIAM & LYSIE
## 754267 NEW HAVEN 52070-014 0853 00200             GUEST CRAIG C
## 754268 NEW HAVEN 52070-014 0853 00300  MEADOWS BERNADETTE (EST)
## 754270 NEW HAVEN 52070-014 0853 00400        KILFEATHER FRANCIS
##                 co_owner          location   mailing_address mailing_city
## 754264 CITY OF NEW HAVEN   75 SOUTH END RD     165 CHURCH ST    NEW HAVEN
## 754265              <NA>  199 SOUTH END RD  199 SOUTH END RD    NEW HAVEN
## 754266              <NA>       11 URIAH ST       11 URIAH ST    NEW HAVEN
## 754267              <NA>  181 SOUTH END RD  181 SOUTH END RD    NEW HAVEN
## 754268              <NA>  169 SOUTH END RD  169 SOUTH END RD    NEW HAVEN
## 754270              <NA> 165 SOUTH END RD  165 SOUTH END RD     NEW HAVEN
##        mailing_state assessed_total assessed_land assessed_building
## 754264            CT         552510        552510                0
## 754265            CT         203490         64330           128240
## 754266            CT         199430         61950           137480
## 754267            CT         161000         67200            93800
## 754268            CT         144550         66150            67900
## 754270            CT         131810         61390            70420
##        pre_year_assessed_total appraised_land appraised_building
## 754264                  552510         789300                 0
## 754265                  203490          91900            183200
## 754266                  199430          88500            196400
## 754267                  161000          96000            134000
## 754268                  144550          94500             97000
## 754270                  131810          87700            100600
##        appraised_outbuilding appraised_extra_feature valuation_year zone
## 754264                     0                      NA           2021  RS2
## 754265                 11900                      NA           2021  RS2
## 754266                     0                      NA           2021  RS2
## 754267                     0                      NA           2021  RS2
## 754268                 15000                      NA           2021  RS2
## 754270                     0                      NA           2021  RS2
##        zone_description model condition condition_description  ayb  eyb
## 754264             <NA>     0      <NA>                 <NA>    0    0
## 754265             <NA>     1         G                 Good 1950 1999
```

```
## 754266             <NA>     1         A               Average 1989 2003
## 754267             <NA>     1         G                  Good 1945 1999
## 754268             <NA>     1         A               Average 1940 1989
## 754270             <NA>     1         A               Average 1940 1989
##        living_area effective_area total_rooms number_of_bedroom number_of_baths
## 754264           0              0          NA                NA              NA
## 754265        1475           1706           6                 3               2
## 754266        1792           1996           6                 3               2
## 754267         864           1129           8                 3               1
## 754268        1040           1289           5                 3               1
## 754270        1080           1238           6                 3               1
##        number_of_half_baths occupancy sale_price           sale_date qualified
## 754264                   NA        NA          0 1975-08-05 20:00:00      <NA>
## 754265                    1         1     132000 2017-01-24 19:00:00         U
## 754266                    1         1     140000 2001-12-30 19:00:00         Q
## 754267                    0         1     183000 2004-07-28 20:00:00         Q
## 754268                    0         1          0 2018-10-03 20:00:00         U
## 754270                    1         1          0 2021-02-24 19:00:00         U
##            prior_sale_date prior_book_page prior_sale_price editor edit_date
## 754264                <NA>            <NA>               NA   <NA>      <NA>
## 754265 2016-05-12 20:00:00       9418/0207                0   <NA>      <NA>
## 754266 1997-10-21 20:00:00       5222/0120                0   <NA>      <NA>
## 754267 1992-02-17 19:00:00       4455/0004            90000   <NA>      <NA>
## 754268 1990-11-01 19:00:00       4307/0304                0   <NA>      <NA>
## 754270 2020-05-26 20:00:00      10005/0104                0   <NA>      <NA>
##        collection_year planning_region state_use state_use_description
## 754264            2023   South Central      <NA>          CITY  MDL-00
## 754265            2023   South Central      1010         Single Family
## 754266            2023   South Central      1010         Single Family
## 754267            2023   South Central      1010         Single Family
## 754268            2023   South Central      1010         Single Family
## 754270            2023   South Central      1010         Single Family
##                                      globalid shape_length shape_area
## 754264 {C0B110A2-1E30-4235-BB9D-D68D554110BA}     890.4312 17083.1649
## 754265 {3F400E85-FA62-43AC-AB66-33CE8B941DC1}     190.1392  1797.0473
## 754266 {DB19BE28-C4AB-4BB7-8A22-BAA81389353E}     181.8368  1303.4148
## 754267 {48912D21-5FB7-4F14-9027-09A16FEB90F3}     315.4156  2986.3171
## 754268 {91C7C939-436A-4AE7-B66D-E4669BB0BD0B}     250.3992  2106.8643
## 754270 {526B9FD9-DDBC-4CCE-BCB9-2D46A4B33769}     136.6323   960.4594
```

### Living Area Restrictions

Set the living area to be above 0 and greater than 200 sq ft. This removes uninhabitable or unreasonably small properties that don't fit the criteria.

```
# Check properties with zero living area
# Checked 75 SOUTH END RD: it's a barren land
head(d[d$living_area == 0 & !is.na(d$living_area), ])
```

```
##        town_name                 link                     owner
## 754264 NEW HAVEN 52070-013 0853 00500 CITY OF NEW HAVEN AIRPORT
## 754273 NEW HAVEN 52070-014 0853 00700        CITY OF NEW HAVEN
## 754274 NEW HAVEN 52070-014 0853 00800 CITY OF NEW HAVEN AIRPORT
```

```
## 754275 NEW HAVEN 52070-014 0853 00900         CITY OF NEW HAVEN
## 754276 NEW HAVEN 52070-014 0853 01000         CITY OF NEW HAVEN
## 754277 NEW HAVEN 52070-014 0853 01100         CITY OF NEW HAVEN
##                co_owner         location mailing_address mailing_city
## 754264 CITY OF NEW HAVEN  75 SOUTH END RD   165 CHURCH ST    NEW HAVEN
## 754273             <NA> 153 SOUTH END RD   165 CHURCH ST    NEW HAVEN
## 754274 CITY OF NEW HAVEN     SOUTH END RD   165 CHURCH ST    NEW HAVEN
## 754275             <NA> 107 SOUTH END RD   165 CHURCH ST    NEW HAVEN
## 754276             <NA> 101 SOUTH END RD   165 CHURCH ST    NEW HAVEN
## 754277             <NA>  95 SOUTH END RD   165 CHURCH ST    NEW HAVEN
##        mailing_state assessed_total assessed_land assessed_building
## 754264            CT         552510        552510                 0
## 754273            CT          64050         64050                 0
## 754274            CT         192710        192710                 0
## 754275            CT          68460         68460                 0
## 754276            CT          59990         59990                 0
## 754277            CT          64540         64540                 0
##        pre_year_assessed_total appraised_land appraised_building
## 754264                  552510         789300                  0
## 754273                   64050          91500                  0
## 754274                  192710         275300                  0
## 754275                   68460          97800                  0
## 754276                   59990          85700                  0
## 754277                   64540          92200                  0
##        appraised_outbuilding appraised_extra_feature valuation_year zone
## 754264                     0                     NA           2021  RS2
## 754273                     0                     NA           2021  RS2
## 754274                     0                     NA           2021  RS2
## 754275                     0                     NA           2021  RS2
## 754276                     0                     NA           2021  RS2
## 754277                     0                     NA           2021  RS2
##        zone_description model condition condition_description ayb eyb
## 754264             <NA>     0      <NA>                 <NA>   0   0
## 754273             <NA>     0      <NA>                 <NA>   0   0
## 754274             <NA>     0      <NA>                 <NA>   0   0
## 754275             <NA>     0      <NA>                 <NA>   0   0
## 754276             <NA>     0      <NA>                 <NA>   0   0
## 754277             <NA>     0      <NA>                 <NA>   0   0
##        living_area effective_area total_rooms number_of_bedroom number_of_baths
## 754264           0              0          NA                NA              NA
## 754273           0              0          NA                NA              NA
## 754274           0              0          NA                NA              NA
## 754275           0              0          NA                NA              NA
## 754276           0              0          NA                NA              NA
## 754277           0              0          NA                NA              NA
##        number_of_half_baths occupancy sale_price           sale_date qualified
## 754264                   NA        NA          0 1975-08-05 20:00:00      <NA>
## 754273                   NA        NA          0 1998-09-01 20:00:00         U
## 754274                   NA        NA          0 1975-08-05 20:00:00      <NA>
## 754275                   NA        NA          0 1967-10-29 19:00:00      <NA>
## 754276                   NA        NA          0 1968-01-03 19:00:00      <NA>
## 754277                   NA        NA          0 1978-09-20 20:00:00      <NA>
##        prior_sale_date prior_book_page prior_sale_price editor edit_date
## 754264            <NA>            <NA>               NA   <NA>      <NA>
```

```
## 754273            <NA>            <NA>               NA   <NA>      <NA>
## 754274            <NA>            <NA>               NA   <NA>      <NA>
## 754275            <NA>            <NA>               NA   <NA>      <NA>
## 754276            <NA>            <NA>               NA   <NA>      <NA>
## 754277            <NA>            <NA>               NA   <NA>      <NA>
##        collection_year planning_region state_use state_use_description
## 754264            2023   South Central      <NA>          CITY  MDL-00
## 754273            2023   South Central      <NA>     MUNICIPAL  MDL-00
## 754274            2023   South Central      <NA>          CITY  MDL-00
## 754275            2023   South Central      <NA>     MUNICIPAL  MDL-00
## 754276            2023   South Central      <NA>     MUNICIPAL  MDL-00
## 754277            2023   South Central      <NA>     MUNICIPAL  MDL-00
##                                      globalid shape_length shape_area
## 754264 {C0B110A2-1E30-4235-BB9D-D68D554110BA}     890.4312 17083.1649
## 754273 {4B0503D5-A371-4C9B-B8E2-FD4ECF38481F}     160.1361  1469.6944
## 754274 {70878141-B1BB-48F6-8E7C-D8748789A4D4}     132.5902   716.3952
## 754275 {2E5F9B42-AD41-42EC-A44C-0BCD53B5F4E0}     237.0975  3485.2293
## 754276 {66245688-7B45-4F93-90E3-EFDF87CCE6D2}     150.0306  1046.8346
## 754277 {B96E5542-FEC6-46D7-90F6-710206D59D06}     190.1728  2143.3247
```

```
sum(d$living_area == 0, na.rm = T)
```

```
## [1] 973
```

```
d <- d[d$living_area != 0 | is.na(d$living_area), ]
head(d)
```

```
##        town_name                 link                      owner co_owner
## 754265 NEW HAVEN 52070-014 0853 00100  PEREZ-RAMIREZ NOE MARTIN     <NA>
## 754266 NEW HAVEN 52070-014 0853 00101 RODRIGUEZ WILLIAM & LYSIE     <NA>
## 754267 NEW HAVEN 52070-014 0853 00200             GUEST CRAIG C     <NA>
## 754268 NEW HAVEN 52070-014 0853 00300  MEADOWS BERNADETTE (EST)     <NA>
## 754270 NEW HAVEN 52070-014 0853 00400        KILFEATHER FRANCIS     <NA>
## 754271 NEW HAVEN 52070-014 0853 00500              MASUD SYED A     <NA>
##                location   mailing_address mailing_city mailing_state
## 754265  199 SOUTH END RD  199 SOUTH END RD    NEW HAVEN            CT
## 754266       11 URIAH ST       11 URIAH ST    NEW HAVEN            CT
## 754267  181 SOUTH END RD  181 SOUTH END RD    NEW HAVEN            CT
## 754268  169 SOUTH END RD  169 SOUTH END RD    NEW HAVEN            CT
## 754270 165 SOUTH END RD  165 SOUTH END RD     NEW HAVEN            CT
## 754271  161 SOUTH END RD 161  SOUTH END RD    NEW HAVEN            CT
##        assessed_total assessed_land assessed_building pre_year_assessed_total
## 754265         203490         64330            128240                  203490
## 754266         199430         61950            137480                  199430
## 754267         161000         67200             93800                  161000
## 754268         144550         66150             67900                  144550
## 754270         131810         61390             70420                  131810
## 754271         133000         59990             73010                  133000
##        appraised_land appraised_building appraised_outbuilding
## 754265          91900             183200                 11900
## 754266          88500             196400                     0
## 754267          96000             134000                     0
## 754268          94500              97000                 15000
```

```
## 754270          87700             100600                    0
## 754271          85700             104300                    0
##        appraised_extra_feature valuation_year zone zone_description model
## 754265                      NA           2021  RS2             <NA>     1
## 754266                      NA           2021  RS2             <NA>     1
## 754267                      NA           2021  RS2             <NA>     1
## 754268                      NA           2021  RS2             <NA>     1
## 754270                      NA           2021  RS2             <NA>     1
## 754271                      NA           2021  RS2             <NA>     1
##        condition condition_description  ayb  eyb living_area effective_area
## 754265         G                 Good 1950 1999        1475           1706
## 754266         A              Average 1989 2003        1792           1996
## 754267         G                 Good 1945 1999         864           1129
## 754268         A              Average 1940 1989        1040           1289
## 754270         A              Average 1940 1989        1080           1238
## 754271         A              Average 1930 1989        1040           1248
##        total_rooms number_of_bedroom number_of_baths number_of_half_baths
## 754265           6                 3               2                    1
## 754266           6                 3               2                    1
## 754267           8                 3               1                    0
## 754268           5                 3               1                    0
## 754270           6                 3               1                    1
## 754271           6                 2               1                    0
##        occupancy sale_price           sale_date qualified     prior_sale_date
## 754265         1     132000 2017-01-24 19:00:00         U 2016-05-12 20:00:00
## 754266         1     140000 2001-12-30 19:00:00         Q 1997-10-21 20:00:00
## 754267         1     183000 2004-07-28 20:00:00         Q 1992-02-17 19:00:00
## 754268         1          0 2018-10-03 20:00:00         U 1990-11-01 19:00:00
## 754270         1          0 2021-02-24 19:00:00         U 2020-05-26 20:00:00
## 754271         1     142000 2011-06-23 20:00:00         Q 2002-03-12 19:00:00
##        prior_book_page prior_sale_price editor edit_date collection_year
## 754265       9418/0207                0   <NA>      <NA>            2023
## 754266       5222/0120                0   <NA>      <NA>            2023
## 754267       4455/0004            90000   <NA>      <NA>            2023
## 754268       4307/0304                0   <NA>      <NA>            2023
## 754270      10005/0104                0   <NA>      <NA>            2023
## 754271       6068/0103            81500   <NA>      <NA>            2023
##        planning_region state_use state_use_description
## 754265   South Central      1010         Single Family
## 754266   South Central      1010         Single Family
## 754267   South Central      1010         Single Family
## 754268   South Central      1010         Single Family
## 754270   South Central      1010         Single Family
## 754271   South Central      1010         Single Family
##                                    globalid shape_length shape_area
## 754265 {3F400E85-FA62-43AC-AB66-33CE8B941DC1}     190.1392  1797.0473
## 754266 {DB19BE28-C4AB-4BB7-8A22-BAA81389353E}     181.8368  1303.4148
## 754267 {48912D21-5FB7-4F14-9027-09A16FEB90F3}     315.4156  2986.3171
## 754268 {91C7C939-436A-4AE7-B66D-E4669BB0BD0B}     250.3992  2106.8643
## 754270 {526B9FD9-DDBC-4CCE-BCB9-2D46A4B33769}     136.6323   960.4594
## 754271 {2E2C8B81-0538-4ACF-BEB3-E813DDC75502}     136.5403   976.3963
```

```r
# Check properties with less than 200 living area
sum(d$living_area < 200, na.rm = T)
```

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

```
# We do not want to live in areas that have less
d <- d[d$living_area > 200 | is.na(d$living_area), ]
head(d)
```

```
##        town_name                 link                     owner co_owner
## 754265 NEW HAVEN 52070-014 0853 00100  PEREZ-RAMIREZ NOE MARTIN     <NA>
## 754266 NEW HAVEN 52070-014 0853 00101 RODRIGUEZ WILLIAM & LYSIE     <NA>
## 754267 NEW HAVEN 52070-014 0853 00200             GUEST CRAIG C     <NA>
## 754268 NEW HAVEN 52070-014 0853 00300  MEADOWS BERNADETTE (EST)     <NA>
## 754270 NEW HAVEN 52070-014 0853 00400        KILFEATHER FRANCIS     <NA>
## 754271 NEW HAVEN 52070-014 0853 00500              MASUD SYED A     <NA>
##                location   mailing_address mailing_city mailing_state
## 754265  199 SOUTH END RD  199 SOUTH END RD    NEW HAVEN            CT
## 754266       11 URIAH ST       11 URIAH ST    NEW HAVEN            CT
## 754267  181 SOUTH END RD  181 SOUTH END RD    NEW HAVEN            CT
## 754268  169 SOUTH END RD  169 SOUTH END RD    NEW HAVEN            CT
## 754270 165 SOUTH END RD  165 SOUTH END RD     NEW HAVEN            CT
## 754271  161 SOUTH END RD 161  SOUTH END RD    NEW HAVEN            CT
##        assessed_total assessed_land assessed_building pre_year_assessed_total
## 754265         203490         64330            128240                  203490
## 754266         199430         61950            137480                  199430
## 754267         161000         67200             93800                  161000
## 754268         144550         66150             67900                  144550
## 754270         131810         61390             70420                  131810
## 754271         133000         59990             73010                  133000
##        appraised_land appraised_building appraised_outbuilding
## 754265          91900             183200                 11900
## 754266          88500             196400                     0
## 754267          96000             134000                     0
## 754268          94500              97000                 15000
## 754270          87700             100600                     0
## 754271          85700             104300                     0
##        appraised_extra_feature valuation_year zone zone_description model
## 754265                      NA           2021  RS2             <NA>     1
## 754266                      NA           2021  RS2             <NA>     1
## 754267                      NA           2021  RS2             <NA>     1
## 754268                      NA           2021  RS2             <NA>     1
## 754270                      NA           2021  RS2             <NA>     1
## 754271                      NA           2021  RS2             <NA>     1
##        condition condition_description  ayb  eyb living_area effective_area
## 754265         G                 Good 1950 1999        1475           1706
## 754266         A              Average 1989 2003        1792           1996
## 754267         G                 Good 1945 1999         864           1129
## 754268         A              Average 1940 1989        1040           1289
## 754270         A              Average 1940 1989        1080           1238
## 754271         A              Average 1930 1989        1040           1248
##        total_rooms number_of_bedroom number_of_baths number_of_half_baths
## 754265           6                 3               2                    1
## 754266           6                 3               2                    1
## 754267           8                 3               1                    0
## 754268           5                 3               1                    0
## 754270           6                 3               1                    1
```

```
## 754271           6                 2               1                    0
##        occupancy sale_price           sale_date qualified     prior_sale_date
## 754265         1     132000 2017-01-24 19:00:00         U 2016-05-12 20:00:00
## 754266         1     140000 2001-12-30 19:00:00         Q 1997-10-21 20:00:00
## 754267         1     183000 2004-07-28 20:00:00         Q 1992-02-17 19:00:00
## 754268         1          0 2018-10-03 20:00:00         U 1990-11-01 19:00:00
## 754270         1          0 2021-02-24 19:00:00         U 2020-05-26 20:00:00
## 754271         1     142000 2011-06-23 20:00:00         Q 2002-03-12 19:00:00
##        prior_book_page prior_sale_price editor edit_date collection_year
## 754265       9418/0207                0   <NA>      <NA>            2023
## 754266       5222/0120                0   <NA>      <NA>            2023
## 754267       4455/0004            90000   <NA>      <NA>            2023
## 754268       4307/0304                0   <NA>      <NA>            2023
## 754270      10005/0104                0   <NA>      <NA>            2023
## 754271       6068/0103            81500   <NA>      <NA>            2023
##        planning_region state_use state_use_description
## 754265   South Central      1010         Single Family
## 754266   South Central      1010         Single Family
## 754267   South Central      1010         Single Family
## 754268   South Central      1010         Single Family
## 754270   South Central      1010         Single Family
## 754271   South Central      1010         Single Family
##                                    globalid shape_length shape_area
## 754265 {3F400E85-FA62-43AC-AB66-33CE8B941DC1}     190.1392  1797.0473
## 754266 {DB19BE28-C4AB-4BB7-8A22-BAA81389353E}     181.8368  1303.4148
## 754267 {48912D21-5FB7-4F14-9027-09A16FEB90F3}     315.4156  2986.3171
## 754268 {91C7C939-436A-4AE7-B66D-E4669BB0BD0B}     250.3992  2106.8643
## 754270 {526B9FD9-DDBC-4CCE-BCB9-2D46A4B33769}     136.6323   960.4594
## 754271 {2E2C8B81-0538-4ACF-BEB3-E813DDC75502}     136.5403   976.3963
```

### Filtering by Town Name

Properties with missing town names often have incomplete data. We verify and remove such entries after checking their attributes.

```
# Check properties with missing town names
nrow(d)
```

```
## [1] 17275
```

```
dnt <- d[is.na(d$town_name), ]
nrow(dnt)
```

```
## [1] 611
```

```
# Inspect columns for properties with missing town names
# These properties are not real! They have no value
table(dnt$assessed_total, useNA = 'always')
```

```
##
## <NA>
##  611
```

```r
table(dnt$number_of_bedroom, useNA = 'always')
```

```
##
## <NA>
##  611
```

```r
table(dnt$number_of_baths, useNA = 'always')
```

```
##
## <NA>
##  611
```

```r
table(dnt$zone, useNA = 'always')
```

```
##
## <NA>
##  611
```

```r
# Remove entries with missing town names
d <- d[!is.na(d$town_name), ]
```

## Main Step (2): Data Cleaning

Note that definitions and categorization of single-family homes vary, so thorough examination is necessary to make reasonable classification decisions.

```r
# `state_use_description` provides finer categorization than `state_use`
length(unique(d$state_use_description))
```

```
## [1] 174
```

```r
length(unique(d$state_use))
```

```
## [1] 94
```

```r
# List all state use descriptions for reference
sort(table(d$state_use_description, useNA = 'always'))
```

```
##
##                 <NA>   ACC COM LD  MDL-00   ACC COM LD  MDL-94
##                    0                   1                   1
##   APT Over12  MDL-01   APT Over12  MDL-96   AUTO V S&S  MDL-96
##                    1                   1                   1
##             CAR WASH           CHILD CARE      CHRCH MDL-02 2F
##                    1                   1                   1
##   CHURCH HSE  MDL-01           CHURCH SCH            CITY LIBR
##                    1                   1                   1
##       CITY MDL-02 2F           ELEC PLANT   EX MED CND  MDL-06
```

```
##                   1                   1                   1
##       EXEMPT  MDL-96   EXEMPT COM  MDL-95   GAS ST SRV  MDL-94
##                   1                   1                   1
##           GAS SUBSTA    HSNG AUTH  MDL-96   INDUSTRIAL  MDL-95
##                   1                   1                   1
##      MARINAS  MDL-00      MARINAS  MDL-01      MARINAS  MDL-94
##                   1                   1                   1
##    MIXED USE  MDL-95   MOVIE THTR  MDL-94   MOVIE THTR  MDL-96
##                   1                   1                   1
##   MUN POLICE  MDL-96      MUNIC MDL-02 2F   NON-PROFIT  MDL-96
##                   1                   1                   1
## NON-PROFIT MDL-02 3F           OFFICE BLD   OFFICE BLD  MDL-95
##                   1                   1                   1
##   PLAZAwAnch  MDL-94             POST OFF   PUB SRV RR  MDL-96
##                   1                   1                   1
##     PVT COLL  MDL-00    PVT COLL  MDL-96   PVT SCHOOL  MDL-00
##                   1                   1                   1
##   PVT SCHOOL  MDL-96    PVT UNIV  MDL-00    PVT UNIV  MDL-01
##                   1                   1                   1
##    R-D FACIL  MDL-96    REC FACIL  MDL-01    RELIGIOUS  MDL-96
##                   1                   1                   1
##  RELIGIOUS MDL-02 2F  RELIGIOUS MDL-02 3F           REST/CLUBS
##                   1                   1                   1
##   STORE/SHOP  MDL-95    TANKS LNG  MDL-94    TANKS LNG  MDL-95
##                   1                   1                   1
##           TENNIS CLB      THEATER  MDL-94      THEATER  MDL-96
##                   1                   1                   1
##             TRK TERM   AUTO S S&S  MDL-95   AUTO S S&S  MDL-96
##                   1                   2                   2
##   CHARITABLE  MDL-01       CHURCH  MDL-96         CITY  MDL-00
##                   2                   2                   2
##         CITY  MDL-95            CITY ADMN            CITY FIRE
##                   2                   2                   2
##   COM BLD NL  MDL-94             COMM IND            EDUC BLDG
##                   2                   2                   2
##   FRATNL ORG  MDL-95             IND BLDG   INDUSTRIAL  MDL-94
##                   2                   2                   2
##           Multi Hses    R-D FACIL  MDL-94    REC FACIL  MDL-96
##                   2                   2                   2
##    STATE DOT  MDL-95      US GOVT  MDL-01           VAC POT BL
##                   2                   2                   2
##           YACHT CLUB   Bording Hs  MDL-94     CEMETERY  MDL-94
##                   2                   3                   3
##       CHURCH  MDL-01   CITY SCHOO  MDL-94    COMM WHSE  MDL-95
##                   3                   3                   3
##       EXEMPT  MDL-01                 GYMS       HOTELS  MDL-94
##                   3                   3                   3
##   IND SHP/GR  MDL-96    IND WHSES  MDL-94             MUN LIBR
##                   3                   3                   3
##   MUN POLICE  MDL-94    MUNICIPAL  MDL-01             PARK LOT
##                   3                   3                   3
##     PVT UNIV  MDL-95             SUPERMKT    TEL X STA  MDL-96
##                   3                   3                   3
##              VAC BLD         CITY  MDL-01             DAY CARE
```

```
##                   3                     4                   4
##     EXEMPT MDL-02 2F    MIXED USE  MDL-96    MUNICIPAL  MDL-95
##                   4                     4                   4
## NON-PROFIT MDL-02 2F   OFFICE BLD  MDL-96     PVT HOSP  MDL-96
##                   4                     4                   4
##    REC FACIL  MDL-94  APT5 - 12 R  MDL-03         CITY  MDL-96
##                   4                     5                   5
##    CLRGY HSE  MDL-01    CLRGY HSE  MDL-94           DEVEL LAND
##                   5                     5                   5
##       EXEMPT  MDL-94   HSG AUTH MDL-02 2F   IND OFFICE  MDL-94
##                   5                     5                   5
##    MIX USE R  MDL-94    MUNICIPAL  MDL-00   NON-PROFIT  MDL-01
##                   5                     5                   5
##   PVT SCHOOL  MDL-01   PVT SCHOOL  MDL-94   STORE/SHOP  MDL-96
##                   5                     5                   5
##      US GOVT  MDL-94   AUTO V S&S  MDL-95   ELECSUBSTA  MDL-96
##                   5                     6                   6
##               MOTELS           NURSING HM    TANKS LNG  MDL-96
##                   6                     6                   6
##   AUTO V S&S  MDL-94    COMM WHSE  MDL-94   GAS ST SRV  MDL-95
##                   7                     7                   7
##    HSNG AUTH  MDL-00         CITY  MDL-94            MUN FIRE
##                   7                     8                   8
##   Mun R Cndo  MDL-05   MUN SCHOOL  MDL-94     PVT UNIV  MDL-96
##                   8                     8                   8
##   CHARITABLE  MDL-94           FUNERAL HM    MIX USE R  MDL-01
##                   9                     9                   9
##     PVT COLL  MDL-94    RELIGIOUS  MDL-01    MUNICIPAL  MDL-96
##                   9                     9                  10
##     PARK GAR  MDL-96            BANK BLDG     PVT HOSP  MDL-94
##                  10                    11                  11
##    MUNICIPAL  MDL-94            PROF BLDG    RELIGIOUS  MDL-94
##                  12                    12                  12
##       SFR Riverfront   IND OFFICE  MDL-96   EXEMPT COM  MDL-96
##                  12                    13                  15
##   FRATNL ORG  MDL-94   RELIG COMM  MDL-94    COMM WHSE  MDL-96
##                  18                    19                  20
##    HSNG AUTH  MDL-94     GAS MART  MDL-94           SFR In-Law
##                  21                    22                  23
##   EXEMPT COM  MDL-94    MIXED USE  MDL-01   IND SHP/GR  MDL-95
##                  25                    26                  33
##    HSNG AUTH  MDL-01            SFR Water   NON-PROFIT  MDL-94
##                  34                    35                  38
##   INDUSTRIAL  MDL-96     PVT UNIV  MDL-94    IND WHSES  MDL-96
##                  61                    75                  77
##   REST/CLUBS  MDL-94   SVC SHP/GA  MDL-95           APT 4-Unit
##                  82                    95                 122
##       CHURCH  MDL-94   OFFICE BLD  MDL-94         Three Family
##                 130                   134                 211
##   APT Over12  MDL-94   STORE/SHOP  MDL-94    APT5 - 12  MDL-94
##                 219                   233                 235
##    MIXED USE  MDL-94           Two Family          Condominium
##                 468                  2530                2703
##        Single Family
```

```
##                 8453
```

```
# Extract the single family homes using the regex exp
state_desc <- tolower(names(table(d$state_use_description, useNA = 'always')))
single_desc <- state_desc[grep("^(1 f|single|sf|one f|s f|sin|res single)",
                               state_desc)]

# Filter single-family properties based on descriptions
c <- d[which(tolower(d$state_use_description) %in% single_desc), ]
table(c$state_use, useNA = 'always')
```

```
##
## 1010 1012 1013 1015 <NA>
## 8453   23   35   12    0
```

```
sort(table(c$state_use, useNA = 'always'))
```

```
##
## <NA> 1015 1012 1013 1010
##    0   12   23   35 8453
```

```
sort(table(c$state_use_description, useNA = 'always'))
```

```
##
##           <NA> SFR Riverfront     SFR In-Law      SFR Water  Single Family
##              0             12             23             35           8453
```

Investigate properties that are NOT classified as single-family. We need to evaluate whether properties like `Condominium` and `Apartment` can be reasonably included as single-family homes.

```
# Filter non-single-family properties for further inspection
ce <- d[!(tolower(d$state_use_description) %in% single_desc), ]
sort(table(ce$state_use))
```

```
##
## 1120 1400 3090 3140 3230  326 3340 3350  340 3500 3620 3640 3750 3840 4040 4220
##    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1    1
## 4280 9028 9059 9092 1090 1310 3165 3222 3310 3510 3841 4022 9018 9025 9027 1300
##    1    1    1    1    2    2    2    2    2    2    2    2    2    2    2    3
## 3000 3240 3370 3760 4050 4300 9026 9031 9035 9062 3520 9094 1111 3900 9000 9040
##    3    3    3    3    3    3    3    3    3    3    4    4    5    5    5    5
## 9066 9070 3010 3040 4210 4240 3160 3300 9020 9032 9033 9039 9045 1070 3550 9041
##    5    5    6    6    6    6    7    7    8    8    8    8    8    9    9    9
## 9100 3410 9051 3420 9030 9090 4020 9029 3530 9053 9080 3250 9200 4000 4010 3260
##    9   11   11   12   12   12   13   15   18   19   21   22   38   61   77   82
## 3320 1110 9060 3400 1050 3220 3030 1040 1020
##   95  122  130  134  211  233  468 2530 2703
```

```
sort(table(ce$state_use_description))
```

```
##
##   ACC COM LD  MDL-00   ACC COM LD  MDL-94   APT Over12  MDL-01
##                    1                    1                    1
##   APT Over12  MDL-96   AUTO V S&S  MDL-96            CAR WASH
##                    1                    1                    1
##           CHILD CARE      CHRCH MDL-02 2F   CHURCH HSE  MDL-01
##                    1                    1                    1
##           CHURCH SCH            CITY LIBR       CITY MDL-02 2F
##                    1                    1                    1
##           ELEC PLANT   EX MED CND  MDL-06       EXEMPT  MDL-96
##                    1                    1                    1
##   EXEMPT COM  MDL-95   GAS ST SRV  MDL-94           GAS SUBSTA
##                    1                    1                    1
##    HSNG AUTH  MDL-96   INDUSTRIAL  MDL-95      MARINAS  MDL-00
##                    1                    1                    1
##      MARINAS  MDL-01      MARINAS  MDL-94    MIXED USE  MDL-95
##                    1                    1                    1
##   MOVIE THTR  MDL-94   MOVIE THTR  MDL-96   MUN POLICE  MDL-96
##                    1                    1                    1
##      MUNIC MDL-02 2F   NON-PROFIT  MDL-96 NON-PROFIT MDL-02 3F
##                    1                    1                    1
##           OFFICE BLD   OFFICE BLD  MDL-95   PLAZAwAnch  MDL-94
##                    1                    1                    1
##             POST OFF   PUB SRV RR  MDL-96     PVT COLL  MDL-00
##                    1                    1                    1
##     PVT COLL  MDL-96   PVT SCHOOL  MDL-00   PVT SCHOOL  MDL-96
##                    1                    1                    1
##     PVT UNIV  MDL-00     PVT UNIV  MDL-01    R-D FACIL  MDL-96
##                    1                    1                    1
##    REC FACIL  MDL-01    RELIGIOUS  MDL-96  RELIGIOUS MDL-02 2F
##                    1                    1                    1
##  RELIGIOUS MDL-02 3F           REST/CLUBS   STORE/SHOP  MDL-95
##                    1                    1                    1
##    TANKS LNG  MDL-94    TANKS LNG  MDL-95           TENNIS CLB
##                    1                    1                    1
##      THEATER  MDL-94      THEATER  MDL-96             TRK TERM
##                    1                    1                    1
##   AUTO S S&S  MDL-95   AUTO S S&S  MDL-96   CHARITABLE  MDL-01
##                    2                    2                    2
##       CHURCH  MDL-96         CITY  MDL-00         CITY  MDL-95
##                    2                    2                    2
##            CITY ADMN            CITY FIRE   COM BLD NL  MDL-94
##                    2                    2                    2
##             COMM IND            EDUC BLDG   FRATNL ORG  MDL-95
##                    2                    2                    2
##             IND BLDG   INDUSTRIAL  MDL-94           Multi Hses
##                    2                    2                    2
##    R-D FACIL  MDL-94    REC FACIL  MDL-96    STATE DOT  MDL-95
##                    2                    2                    2
##      US GOVT  MDL-01           VAC POT BL           YACHT CLUB
##                    2                    2                    2
##   Bording Hs  MDL-94     CEMETERY  MDL-94       CHURCH  MDL-01
##                    3                    3                    3
##   CITY SCHOO  MDL-94    COMM WHSE  MDL-95       EXEMPT  MDL-01
```

```
##                   3                   3                   3
##                GYMS       HOTELS  MDL-94   IND SHP/GR  MDL-96
##                   3                   3                   3
##    IND WHSES  MDL-94            MUN LIBR   MUN POLICE  MDL-94
##                   3                   3                   3
##    MUNICIPAL  MDL-01            PARK LOT     PVT UNIV  MDL-95
##                   3                   3                   3
##            SUPERMKT    TEL X STA  MDL-96             VAC BLD
##                   3                   3                   3
##         CITY  MDL-01            DAY CARE     EXEMPT MDL-02 2F
##                   4                   4                   4
##    MIXED USE  MDL-96    MUNICIPAL  MDL-95 NON-PROFIT MDL-02 2F
##                   4                   4                   4
##   OFFICE BLD  MDL-96     PVT HOSP  MDL-96    REC FACIL  MDL-94
##                   4                   4                   4
##  APT5 - 12 R  MDL-03         CITY  MDL-96    CLRGY HSE  MDL-01
##                   5                   5                   5
##    CLRGY HSE  MDL-94          DEVEL LAND       EXEMPT  MDL-94
##                   5                   5                   5
##   HSG AUTH MDL-02 2F   IND OFFICE  MDL-94    MIX USE R  MDL-94
##                   5                   5                   5
##    MUNICIPAL  MDL-00   NON-PROFIT  MDL-01   PVT SCHOOL  MDL-01
##                   5                   5                   5
##   PVT SCHOOL  MDL-94   STORE/SHOP  MDL-96      US GOVT  MDL-94
##                   5                   5                   5
##   AUTO V S&S  MDL-95   ELECSUBSTA  MDL-96               MOTELS
##                   6                   6                   6
##           NURSING HM    TANKS LNG  MDL-96   AUTO V S&S  MDL-94
##                   6                   6                   7
##    COMM WHSE  MDL-94   GAS ST SRV  MDL-95    HSNG AUTH  MDL-00
##                   7                   7                   7
##         CITY  MDL-94            MUN FIRE   Mun R Cndo  MDL-05
##                   8                   8                   8
##   MUN SCHOOL  MDL-94     PVT UNIV  MDL-96   CHARITABLE  MDL-94
##                   8                   8                   9
##           FUNERAL HM    MIX USE R  MDL-01     PVT COLL  MDL-94
##                   9                   9                   9
##    RELIGIOUS  MDL-01    MUNICIPAL  MDL-96     PARK GAR  MDL-96
##                   9                  10                  10
##            BANK BLDG     PVT HOSP  MDL-94    MUNICIPAL  MDL-94
##                  11                  11                  12
##            PROF BLDG    RELIGIOUS  MDL-94   IND OFFICE  MDL-96
##                  12                  12                  13
##   EXEMPT COM  MDL-96   FRATNL ORG  MDL-94   RELIG COMM  MDL-94
##                  15                  18                  19
##    COMM WHSE  MDL-96    HSNG AUTH  MDL-94     GAS MART  MDL-94
##                  20                  21                  22
##   EXEMPT COM  MDL-94    MIXED USE  MDL-01   IND SHP/GR  MDL-95
##                  25                  26                  33
##    HSNG AUTH  MDL-01   NON-PROFIT  MDL-94   INDUSTRIAL  MDL-96
##                  34                  38                  61
##     PVT UNIV  MDL-94    IND WHSES  MDL-96   REST/CLUBS  MDL-94
##                  75                  77                  82
##   SVC SHP/GA  MDL-95          APT 4-Unit       CHURCH  MDL-94
```

```
##                   95                  122                  130
##   OFFICE BLD  MDL-94         Three Family   APT Over12  MDL-94
##                  134                  211                  219
##   STORE/SHOP  MDL-94    APT5 - 12  MDL-94    MIXED USE  MDL-94
##                  233                  235                  468
##           Two Family          Condominium
##                 2530                 2703
```

Inspect properties classified as `Condominium` or `Apartment`. These are attractive to single-family buyers, so they are manually checked.

```
# Check properties classified as apartments
ce_apt <- ce[ce$state_use_description %in% c("APT 4-Unit",
                                             "APT5 - 12  MDL-94"), ]

# checked 490 WOODWARD AV: a great 2 bed - 1 bath apartment
# checked 43 HARRINGTON AV: a great 2 bed - 1 bath apartment
# checked 165 MAIN ST ANNEX: a clean apartment!
# checked  29 TERRACE ST: a great 2 bed - 1.5 bath apartments
# checked 70 HARRINGTON AV: a clean apartment!
head(ce_apt[!is.na(ce_apt$state_use_description), ])
```

```
##        town_name                 link                        owner
## 755983 NEW HAVEN 52070-053 0950 01000                   HALL JOHN D
## 756078 NEW HAVEN 52070-054 0933 01200 SABELLICO ALEXANDER & CORSO  *
## 756215 NEW HAVEN 52070-056 0943 01700   HARRINGTON AVENUE PROPERTIES
## 756276 NEW HAVEN 52070-057 0941 03400             LICATA NICHOLAS J
## 756305 NEW HAVEN 52070-057 0942 00101          29 TERRACE STREET LLC
## 756325 NEW HAVEN 52070-057 0942 02300    70-76 HARRINGTON AVENUE LLC
##                       co_owner          location      mailing_address
## 755983                    <NA>   490 WOODWARD AV       19 MARIETTA ST
## 756078 DEBORAH & SHORE ROSEANN   543 WOODWARD AV           PO BOX 825
## 756215                     LLC  43 HARRINGTON AV 3258 COUNTRY CLUB RD
## 756276                    <NA> 165 MAIN ST ANNEX    165 MAIN ST ANNEX
## 756305                    <NA>     29 TERRACE ST     125 UNDERHILL ST
## 756325                    <NA>  70 HARRINGTON AV     70 HARRINGTON AV
##        mailing_city mailing_state assessed_total assessed_land
## 755983       HAMDEN            CT         352520         65940
## 756078       ORANGE            CT         652400         70840
## 756215        BRONX            NY         252140         54320
## 756276    NEW HAVEN            CT         350840         55650
## 756305      YONKERS            NY         277690         56700
## 756325    NEW HAVEN            CT         331870         56700
##        assessed_building pre_year_assessed_total appraised_land
## 755983            276500                 352520          94200
## 756078            577220                 652400         101200
## 756215            197820                 252140          77600
## 756276            295190                 350840          79500
## 756305            220990                 277690          81000
## 756325            275170                 331870          81000
##        appraised_building appraised_outbuilding appraised_extra_feature
## 755983             395000                14400                      NA
## 756078             824600                 6200                      NA
```

```
## 756215             282600                    0                     NA
## 756276             421700                    0                     NA
## 756305             315700                    0                     NA
## 756325             393100                    0                     NA
##        valuation_year   zone zone_description model condition
## 755983           2021    RM1             <NA>    94         A
## 756078           2021 PDD 52             <NA>    94         A
## 756215           2021    RM1             <NA>     2         G
## 756276           2021     BB             <NA>    94         G
## 756305           2021    RM1             <NA>     2         A
## 756325           2021    RM1             <NA>     2         A
##        condition_description  ayb  eyb living_area effective_area total_rooms
## 755983               Average 1966 1981        6630           7435          NA
## 756078               Average 1973 1986        7956           8288          NA
## 756215                  Good 1965 1999        3648           4019          12
## 756276                  Good 1900 1986        3327           3675          NA
## 756305               Average 1968 1989        4256           4708          12
## 756325               Average 1982 2001        4292           4295          16
##        number_of_bedroom number_of_baths number_of_half_baths occupancy
## 755983                NA              NA                   NA         7
## 756078                NA              NA                   NA        10
## 756215                 4               4                    0         4
## 756276                NA              NA                   NA         5
## 756305                 4               4                    0         4
## 756325                 4               4                    4         4
##        sale_price           sale_date qualified     prior_sale_date
## 755983     300000 2016-08-18 20:00:00         Q 1988-12-28 19:00:00
## 756078          0 2008-06-29 20:00:00         U 1988-09-27 20:00:00
## 756215          0 2018-01-25 19:00:00         U 2016-09-27 20:00:00
## 756276     200000 2014-03-17 20:00:00         Q 2012-06-11 20:00:00
## 756305     350000 2015-07-22 20:00:00         U 2009-07-27 20:00:00
## 756325          0 2015-11-29 19:00:00         U 2014-08-12 20:00:00
##        prior_book_page prior_sale_price editor edit_date collection_year
## 755983       4022/0298            82000   <NA>      <NA>            2023
## 756078       3972/0239                0   <NA>      <NA>            2023
## 756215       9478/0109           196500   <NA>      <NA>            2023
## 756276       8842/0239                0   <NA>      <NA>            2023
## 756305       8416/0271           234900   <NA>      <NA>            2023
## 756325       9183/0254           350000   <NA>      <NA>            2023
##        planning_region state_use state_use_description
## 755983   South Central      <NA>     APT5 - 12  MDL-94
## 756078   South Central      <NA>     APT5 - 12  MDL-94
## 756215   South Central      1110            APT 4-Unit
## 756276   South Central      <NA>     APT5 - 12  MDL-94
## 756305   South Central      1110            APT 4-Unit
## 756325   South Central      1110            APT 4-Unit
##                                      globalid shape_length shape_area
## 755983 {0181EC9E-9BB6-450E-AC4D-0210391AEB29}     299.3853   3390.332
## 756078 {4D03AC23-2537-4B6F-A9FA-97035B99C44F}     374.1220   4862.132
## 756215 {414D2C27-A210-4530-A907-686FFD5E3BD8}     153.2533   1464.290
## 756276 {FB17FC8E-DA99-4CB0-B366-F5DC1C3512DA}     152.7930   1146.139
## 756305 {281726EA-8D51-4DAA-81BA-451C5DDE3F8C}     182.7052   2063.998
## 756325 {864B1B84-9D02-4F1E-91E7-582BCA7AE510}     183.3375   2073.711
```

```r
# save the description labels for apartments
apt_desc <- state_desc[grep("^apt(?!t)", state_desc, perl = TRUE)]

# Confirm that apartments are legitimate properties
# checked 21 COLBY #CT: it's a great apartment!
# checked 145 COOPER PL: it's a great apartment
tail(d[tolower(d$state_use_description) %in% apt_desc, ])
```

```
##        town_name                 link                  owner co_owner
## 780600 NEW HAVEN 52070-437 1275 01400 H AND R INVESTMENT LLC     <NA>
## 780619 NEW HAVEN 52070-437 1275 01200 H AND R INVESTMENT LLC     <NA>
## 780625 NEW HAVEN 52070-437 1234 00100   IMPERIAL GARDENS LLC     <NA>
## 780978 NEW HAVEN 52070-035 0846 00100        CORA STREET LLC     <NA>
## 781197 NEW HAVEN 52070-438 1226 00100 WEST GATE VENTURES LLC     <NA>
## 781250 NEW HAVEN 52070-438 1226 00101 WEST GATE VENTURES LLC     <NA>
##              location     mailing_address mailing_city mailing_state
## 780600    21 COLBY #CT          19 HOWE ST    NEW HAVEN            CT
## 780619     1 COLBY #CT          19 HOWE ST    NEW HAVEN            CT
## 780625 492 FOUNTAIN ST 167 KINGFISHER LANE    WESTBROOK            CT
## 780978       1 CORA ST      12 OLD TOWN RD      SEYMOUR            CT
## 781197    35 COOPER PL          PO BOX 868     LAKEWOOD            NJ
## 781250   145 COOPER PL          PO BOX 868     LAKEWOOD            NJ
##        assessed_total assessed_land assessed_building pre_year_assessed_total
## 780600         510930         62860            447860                  510930
## 780619         345520         62090            274540                  345520
## 780625        6221670       1260000           4922680                 6221670
## 780978         713090        191450            504910                  713090
## 781197        6364960       1225000           5114130                 6364960
## 781250        5594890       1172500           4368280                 5594890
##        appraised_land appraised_building appraised_outbuilding
## 780600          89800             639800                   300
## 780619          88700             392200                 12700
## 780625        1800000            7032400                 55700
## 780978         273500             721300                 10800
## 781197        1750000            7305900                 36900
## 781250        1675000            6240400                 77300
##        appraised_extra_feature valuation_year    zone zone_description model
## 780600                      NA           2021     RM1             <NA>    94
## 780619                      NA           2021     RM1             <NA>    94
## 780625                      NA           2021 RM1/RS2             <NA>    94
## 780978                      NA           2021     RS2             <NA>    94
## 781197                      NA           2021     RM1             <NA>    94
## 781250                      NA           2021     RM1             <NA>    94
##        condition condition_description  ayb  eyb living_area effective_area
## 780600         A               Average 1950 1981        7862           8713
## 780619         A               Average 1950 1981        4368           4859
## 780625         G                  Good 1960 1986        6626           6830
## 780978         A               Average 1956 1981        7650           8242
## 781197         G                  Good 1947 1986       19305          21385
## 781250         G                  Good 1947 1986        6371           7064
##        total_rooms number_of_bedroom number_of_baths number_of_half_baths
## 780600          NA                NA              NA                   NA
## 780619          NA                NA              NA                   NA
```

```
## 780625          NA                NA              NA                   NA
## 780978          NA                NA              NA                   NA
## 781197          NA                NA              NA                   NA
## 781250          NA                NA              NA                   NA
##        occupancy sale_price           sale_date qualified     prior_sale_date
## 780600         9          0 2020-07-12 20:00:00         U 2009-08-20 20:00:00
## 780619         5          0 2020-07-12 20:00:00         U 2009-08-20 20:00:00
## 780625         6    4250000 2011-08-23 20:00:00         Q 2006-04-26 20:00:00
## 780978         8          0 2018-04-22 20:00:00         U 2000-02-06 19:00:00
## 781197        20   21000000 2021-01-27 19:00:00         U 2012-12-19 19:00:00
## 781250         8   21000000 2021-01-27 19:00:00         U 2012-12-19 19:00:00
##        prior_book_page prior_sale_price editor edit_date collection_year
## 780600       8429/0164                0   <NA>      <NA>            2023
## 780619       8429/0164                0   <NA>      <NA>            2023
## 780625       7569/0036                0   <NA>      <NA>            2023
## 780978       5623/0159           510000   <NA>      <NA>            2023
## 781197       8923/0108         12250000   <NA>      <NA>            2023
## 781250       8923/0108         12250000   <NA>      <NA>            2023
##        planning_region state_use state_use_description
## 780600   South Central      <NA>     APT5 - 12  MDL-94
## 780619   South Central      <NA>     APT5 - 12  MDL-94
## 780625   South Central      <NA>    APT Over12  MDL-94
## 780978   South Central      <NA>     APT5 - 12  MDL-94
## 781197   South Central      <NA>    APT Over12  MDL-94
## 781250   South Central      <NA>    APT Over12  MDL-94
##                                    globalid shape_length shape_area
## 780600 {EC6E0C76-C88C-4212-90FC-6D3CD9D47C47}     217.8941   2124.672
## 780619 {F111046A-322E-44E3-AD87-9BA7BA4581A7}     297.9558   4425.623
## 780625 {025FFD6B-C400-42B2-A0B2-B380D6658485}     816.9821  22992.242
## 780978 {8E65BCDC-1C70-48F6-BBB9-D3441D1C8082}     273.6358   4275.485
## 781197 {43CB8BC4-7BD4-4E80-8DDB-0A31CB8436CA}    1208.8390  33621.909
## 781250 {FCA39552-08BA-4EDA-8C13-9E61E3435B8F}    1775.4764  37970.667
```

```
# Check Condominium

# checked 675 TOWNSEND AV #116: a great 2 bed - 2 bath condominium
# checked 418 WOODWARD AV #26: a great 2 bed - 1 bath condominium
# checked 307 SAINT RONAN ST #A-7: a great 3 bed - 2 bath condominium
# (have walked across this property many times, many people at Yale
# would definitely consider living in this place)
ce_cond <- ce[ce$state_use_description %in% c("Condominium"), ]
head(ce_cond[!is.na(ce_cond$state_use_description), ])
```

```
##        town_name                 link                owner co_owner
## 754398 NEW HAVEN 52070-021 0920 03201   CELENTANO ROBERT J     <NA>
## 754399 NEW HAVEN 52070-021 0920 03202 ANDERSON STEPHANIE L     <NA>
## 754400 NEW HAVEN 52070-021 0920 03203         RAK MONICA S     <NA>
## 754401 NEW HAVEN 52070-021 0920 03204       CARFORA PAMELA     <NA>
## 754402 NEW HAVEN 52070-021 0920 03205  KEBABIAN MARCELLA J     <NA>
## 754403 NEW HAVEN 52070-021 0920 03206         BERT KATHRYN     <NA>
##                     location           mailing_address mailing_city
## 754398 675 TOWNSEND AV #116    675 TOWNSEND AV    116    NEW HAVEN
## 754399 675 TOWNSEND AV #117      675 TOWNSEND AV #117    NEW HAVEN
## 754400 675 TOWNSEND AV #118 675 TOWNSEND AV  UNIT 118    NEW HAVEN
```

```
## 754401 675 TOWNSEND AV #119      675 TOWNSEND AV #119    NEW HAVEN
## 754402 675 TOWNSEND AV #120     675 TOWNSEND AV   120    NEW HAVEN
## 754403 675 TOWNSEND AV #121      675 TOWNSEND AV #121    NEW HAVEN
##        mailing_state assessed_total assessed_land assessed_building
## 754398            CT         146930             0            146930
## 754399            CT         163870             0            161000
## 754400            CT         163870             0            161000
## 754401            CT         126140             0            126140
## 754402            CT         130760             0            130760
## 754403            CT         163870             0            161000
##        pre_year_assessed_total appraised_land appraised_building
## 754398                  146930             0             209900
## 754399                  163870             0             230000
## 754400                  163870             0             230000
## 754401                  126140             0             180200
## 754402                  130760             0             186800
## 754403                  163870             0             230000
##        appraised_outbuilding appraised_extra_feature valuation_year   zone
## 754398                     0                      NA           2021 PDD 48
## 754399                     0                      NA           2021 PDD 48
## 754400                     0                      NA           2021 PDD 48
## 754401                     0                      NA           2021 PDD 48
## 754402                     0                      NA           2021 PDD 48
## 754403                     0                      NA           2021 PDD 48
##        zone_description model condition condition_description  ayb  eyb
## 754398             <NA>     5         G                 Good 1985 2009
## 754399             <NA>     5         G                 Good 1985 2009
## 754400             <NA>     5         G                 Good 1985 2009
## 754401             <NA>     5         A              Average 1985 2005
## 754402             <NA>     5         G                 Good 1985 2009
## 754403             <NA>     5         G                 Good 1985 2009
##        living_area effective_area total_rooms number_of_bedroom number_of_baths
## 754398        1250           1346           5                 2               2
## 754399        1406           1502           5                 2               2
## 754400        1406           1502           5                 2               2
## 754401        1084           1180           4                 2               1
## 754402        1084           1180           4                 2               1
## 754403        1406           1502           5                 2               2
##        number_of_half_baths occupancy sale_price           sale_date qualified
## 754398                    0         1     211500 2006-03-06 19:00:00         U
## 754399                    1         1     286500 2021-11-22 19:00:00         U
## 754400                    1         1     125000 2000-08-30 20:00:00         Q
## 754401                    1         1     165000 2021-12-05 19:00:00         Q
## 754402                    1         1          0 2006-03-27 19:00:00         U
## 754403                    1         1     281500 2004-06-02 20:00:00         Q
##            prior_sale_date prior_book_page prior_sale_price editor edit_date
## 754398 2006-01-17 19:00:00       7467/0267                0   <NA>      <NA>
## 754399 2020-07-19 20:00:00      10028/0291           162686   <NA>      <NA>
## 754400 1987-04-06 20:00:00       3651/0205           175000   <NA>      <NA>
## 754401 2009-02-25 19:00:00       8348/0193           167000   <NA>      <NA>
## 754402 2006-03-27 19:00:00       7536/0059                0   <NA>      <NA>
## 754403 2000-08-22 20:00:00       5718/0049           145000   <NA>      <NA>
##        collection_year planning_region state_use state_use_description
## 754398            2023   South Central      1020           Condominium
```

```
## 754399            2023   South Central      1020           Condominium
## 754400            2023   South Central      1020           Condominium
## 754401            2023   South Central      1020           Condominium
## 754402            2023   South Central      1020           Condominium
## 754403            2023   South Central      1020           Condominium
##                                     globalid shape_length shape_area
## 754398 {C1FF1CA7-FAEF-4C58-ABC3-8E27686135C6}     1436.312   123909.6
## 754399 {F65AC694-2545-4C1E-9AAB-2C7419C3612C}     1437.908   123915.0
## 754400 {BF2E0C7E-62D1-44CC-9603-8686556DD450}     1437.212   124310.8
## 754401 {24322769-C586-461C-B49D-F3A4E6FD4936}     1437.212   124310.8
## 754402 {AAC33810-BC1C-4CBD-B8F8-C57D4A958B8B}     1437.212   124310.8
## 754403 {ABEE1B43-2966-468E-8855-AB2570B2C991}     1437.212   124310.8
```

```
tail(ce_cond[!is.na(ce_cond$state_use_description), ])
```

```
##        town_name                link                         owner
## 781127 NEW HAVEN 52070-042 0950 00223                  DUBNO LUBA C
## 781128 NEW HAVEN 52070-042 0950 00224                  CASEY KEVIN B
## 781129 NEW HAVEN 52070-042 0950 00225            VAZQUEZ JOSEFINA B
## 781130 NEW HAVEN 52070-042 0950 00226 DACOSTA ANTONIO A & ALPHA L &
## 781131 NEW HAVEN 52070-042 0950 00227           MAHONEY PATRICIA M &
## 781132 NEW HAVEN 52070-042 0950 00228      PIETROSIMONE MICHAEL EXE
##                  co_owner            location         mailing_address
## 781127              <NA> 418 WOODWARD AV #23    418 WOODWARD AV U-23
## 781128              <NA> 418 WOODWARD AV #24 418 WOODWARD AV UNIT 24
## 781129              <NA> 418 WOODWARD AV #25      418 WOODWARD AV#25
## 781130              SURV 418 WOODWARD AV #26 418 WOODWARD AV UNIT 26
## 781131 SULLIVAN JOHN PAUL 418 WOODWARD AV #27      418 WOODWARD AV 27
## 781132    ANASTASIO LINDA 418 WOODWARD AV #28      113 LAURELBROOK DR
##        mailing_city mailing_state assessed_total assessed_land
## 781127    NEW HAVEN            CT          77560             0
## 781128    NEW HAVEN            CT          77560             0
## 781129    NEW HAVEN            CT          85260             0
## 781130    NEW HAVEN            CT          83440             0
## 781131    NEW HAVEN            CT          77560             0
## 781132     GUILFORD            CT          77560             0
##        assessed_building pre_year_assessed_total appraised_land
## 781127             77560                  77560              0
## 781128             77560                  77560              0
## 781129             85260                  85260              0
## 781130             83440                  83440              0
## 781131             77560                  77560              0
## 781132             77560                  77560              0
##        appraised_building appraised_outbuilding appraised_extra_feature
## 781127             110800                    0                      NA
## 781128             110800                    0                      NA
## 781129             121800                    0                      NA
## 781130             119200                    0                      NA
## 781131             110800                    0                      NA
## 781132             110800                    0                      NA
##        valuation_year zone zone_description model condition
## 781127           2021  RM1             <NA>     5         A
## 781128           2021  RM1             <NA>     5         A
## 781129           2021  RM1             <NA>     5         A
```

```
## 781130           2021  RM1             <NA>     5         A
## 781131           2021  RM1             <NA>     5         A
## 781132           2021  RM1             <NA>     5         A
##        condition_description  ayb  eyb living_area effective_area total_rooms
## 781127               Average 1986 2005         977            977           4
## 781128               Average 1986 2005         977            977           4
## 781129               Average 1986 2005        1166           1166           4
## 781130               Average 1986 2005        1130           1130           4
## 781131               Average 1985 2005         977            977           4
## 781132               Average 1986 2005         977            977           4
##        number_of_bedroom number_of_baths number_of_half_baths occupancy
## 781127                 2               1                    1         1
## 781128                 2               1                    1         1
## 781129                 2               1                    0         1
## 781130                 2               1                    0         1
## 781131                 2               1                    1         1
## 781132                 2               1                    1         1
##        sale_price           sale_date qualified     prior_sale_date
## 781127     117000 1987-08-20 20:00:00         U                <NA>
## 781128      96500 2018-06-07 20:00:00         Q 2007-05-24 20:00:00
## 781129      63000 1994-10-02 20:00:00         Q 1986-04-27 20:00:00
## 781130          0 1993-05-25 20:00:00         U                <NA>
## 781131          0 2018-02-21 19:00:00         U 1986-04-28 20:00:00
## 781132          0 2023-04-19 20:00:00         U 1998-05-28 20:00:00
##        prior_book_page prior_sale_price editor edit_date collection_year
## 781127            <NA>               NA   <NA>      <NA>            2023
## 781128       7965/0056           165000   <NA>      <NA>            2023
## 781129       3445/0113                0   <NA>      <NA>            2023
## 781130            <NA>               NA   <NA>      <NA>            2023
## 781131       3445/0248                0   <NA>      <NA>            2023
## 781132       5317/0333            51000   <NA>      <NA>            2023
##        planning_region state_use state_use_description
## 781127   South Central      1020           Condominium
## 781128   South Central      1020           Condominium
## 781129   South Central      1020           Condominium
## 781130   South Central      1020           Condominium
## 781131   South Central      1020           Condominium
## 781132   South Central      1020           Condominium
##                                      globalid shape_length shape_area
## 781127 {0D9FFBDB-339A-4586-99BC-DF75B78F086F}     483.8936   12509.72
## 781128 {6EB553F3-9ECD-475F-9358-DA65DB1D90B4}     483.8936   12509.72
## 781129 {BF991809-1D3F-491D-BB28-5FD7EA57454B}     483.8936   12509.72
## 781130 {3DB64A53-C31B-41A6-8009-AC58703A47F5}     483.8936   12509.72
## 781131 {F65B7056-4447-4DA2-A762-EFB2E3CF3DE6}     483.8936   12509.72
## 781132 {3B70C5E0-4DC3-403C-A010-06E69939A886}     483.8936   12509.72
```

```
head(tail(ce_cond[!is.na(ce_cond$state_use_description), ], 1000))
```

```
##        town_name                link                          owner
## 767405 NEW HAVEN 52070-220 0410 01006            GORDON JOHN S TTEE
## 767406 NEW HAVEN 52070-220 0410 01007         WELCH G HAROLD JR TTEE
## 767407 NEW HAVEN 52070-220 0410 01008               WISNER MELISSA A
## 767408 NEW HAVEN 52070-220 0410 01009 POGGE THOMAS WINFRIED MENKO &
## 767409 NEW HAVEN 52070-220 0410 01010   COOKE EDWARD & WARNER CAROL
```

```
## 767410 NEW HAVEN 52070-220 0410 01011                 ZHONG WEIMIN
##                             co_owner               location
## 767405                          <NA> 309 SAINT RONAN ST #A-6
## 767406 G HAROLD WELCH JR REVOCABLE TR 307 SAINT RONAN ST #A-7
## 767407                          <NA> 313 SAINT RONAN ST #B-1
## 767408               TONG LYNN LING 313 SAINT RONAN ST #B-2
## 767409                          <NA> 311 SAINT RONAN ST #B-3
## 767410                          <NA> 311 SAINT RONAN ST #B-4
##               mailing_address mailing_city mailing_state assessed_total
## 767405      309 SAINT RONAN ST    NEW HAVEN            CT         354340
## 767406        257 HICKORY LANE    BETHLEHEM            CT         431480
## 767407 313 SAINT RONAN ST #B-1    NEW HAVEN            CT         387590
## 767408 313 SAINT RONAN ST  B-2    NEW HAVEN            CT         254310
## 767409 311 SAINT RONAN ST #B-3    NEW HAVEN            CT         377440
## 767410 311 SAINT RONAN ST #B-4    NEW HAVEN            CT         350700
##        assessed_land assessed_building pre_year_assessed_total appraised_land
## 767405             0            351680                  354340             0
## 767406             0            428820                  431480             0
## 767407             0            385280                  387590             0
## 767408             0            251650                  254310             0
## 767409             0            377440                  377440             0
## 767410             0            348040                  350700             0
##        appraised_building appraised_outbuilding appraised_extra_feature
## 767405             502400                     0                      NA
## 767406             612600                     0                      NA
## 767407             550400                     0                      NA
## 767408             359500                     0                      NA
## 767409             539200                     0                      NA
## 767410             497200                     0                      NA
##        valuation_year zone zone_description model condition
## 767405           2021  RS1             <NA>     5         G
## 767406           2021  RS1             <NA>     5         G
## 767407           2021  RS1             <NA>     5         A
## 767408           2021  RS1             <NA>     5         G
## 767409           2021  RS1             <NA>     5         A
## 767410           2021  RS1             <NA>     5         G
##        condition_description  ayb  eyb living_area effective_area total_rooms
## 767405                  Good 1926 2001        1277           1277           5
## 767406                  Good 1926 2001        1641           1641           6
## 767407               Average 1926 1991        1706           1706           6
## 767408                  Good 1926 2001        1230           1230           5
## 767409               Average 1921 1991        1653           1653           6
## 767410                  Good 1926 2001        1241           1241           5
##        number_of_bedroom number_of_baths number_of_half_baths occupancy
## 767405                 2               1                    0         1
## 767406                 3               2                    0         1
## 767407                 3               2                    0         1
## 767408                 2               1                    0         1
## 767409                 2               2                    0         1
## 767410                 2               2                    0         1
##        sale_price           sale_date qualified     prior_sale_date
## 767405     310000 2009-06-15 20:00:00         U 2009-06-15 20:00:00
## 767406          0 2016-05-24 20:00:00         U 2009-05-14 20:00:00
## 767407     773500 2022-06-14 20:00:00         Q 2012-04-08 20:00:00
```

```
## 767408     420000 2008-07-01 20:00:00         Q 1998-01-23 19:00:00
## 767409     335000 2014-04-01 20:00:00         Q 1983-02-28 19:00:00
## 767410     175000 2013-10-20 20:00:00         Q 1973-09-09 20:00:00
##        prior_book_page prior_sale_price editor edit_date collection_year
## 767405       8397/0170                0   <NA>      <NA>            2023
## 767406       8383/0162                0   <NA>      <NA>            2023
## 767407       8816/0028                0   <NA>      <NA>            2023
## 767408       5264/0331            95000   <NA>      <NA>            2023
## 767409       3041/0330            45900   <NA>      <NA>            2023
## 767410             0/0            45000   <NA>      <NA>            2023
##        planning_region state_use state_use_description
## 767405   South Central      1020           Condominium
## 767406   South Central      1020           Condominium
## 767407   South Central      1020           Condominium
## 767408   South Central      1020           Condominium
## 767409   South Central      1020           Condominium
## 767410   South Central      1020           Condominium
##                                      globalid shape_length shape_area
## 767405 {57FE1430-25AF-45CF-88C3-2643EFF6F651}     370.9404   7937.146
## 767406 {D7E1C75D-607C-4BC3-A3FF-90AFD173781C}     370.9404   7937.146
## 767407 {6754858D-999B-40CD-9433-907F9FCAEB4E}     370.9404   7937.146
## 767408 {A66667B4-D368-423F-8DE8-A7F7CB21CC9A}     370.9404   7937.146
## 767409 {F0DB0F4E-8D02-40F8-948C-A022BE0B6180}     370.9404   7937.146
## 767410 {5815A754-46A2-4C31-AA53-990473ED0BDC}     370.9404   7937.146
```

```
# Save descriptions for condominiums
condo_desc <- state_desc[grep("^(condo)", state_desc, perl = TRUE)]

# Confirm that condominiums are legitimate propertiess
tail(d[tolower(d$state_use_description) %in% condo_desc, ])
```

```
##        town_name                 link                        owner
## 781127 NEW HAVEN 52070-042 0950 00223                DUBNO LUBA C
## 781128 NEW HAVEN 52070-042 0950 00224                CASEY KEVIN B
## 781129 NEW HAVEN 52070-042 0950 00225           VAZQUEZ JOSEFINA B
## 781130 NEW HAVEN 52070-042 0950 00226 DACOSTA ANTONIO A & ALPHA L &
## 781131 NEW HAVEN 52070-042 0950 00227          MAHONEY PATRICIA M &
## 781132 NEW HAVEN 52070-042 0950 00228      PIETROSIMONE MICHAEL EXE
##                  co_owner           location         mailing_address
## 781127              <NA> 418 WOODWARD AV #23    418 WOODWARD AV U-23
## 781128              <NA> 418 WOODWARD AV #24 418 WOODWARD AV UNIT 24
## 781129              <NA> 418 WOODWARD AV #25      418 WOODWARD AV#25
## 781130              SURV 418 WOODWARD AV #26 418 WOODWARD AV UNIT 26
## 781131 SULLIVAN JOHN PAUL 418 WOODWARD AV #27      418 WOODWARD AV 27
## 781132    ANASTASIO LINDA 418 WOODWARD AV #28      113 LAURELBROOK DR
##        mailing_city mailing_state assessed_total assessed_land
## 781127    NEW HAVEN            CT          77560             0
## 781128    NEW HAVEN            CT          77560             0
## 781129    NEW HAVEN            CT          85260             0
## 781130    NEW HAVEN            CT          83440             0
## 781131    NEW HAVEN            CT          77560             0
## 781132     GUILFORD            CT          77560             0
##        assessed_building pre_year_assessed_total appraised_land
## 781127             77560                  77560              0
```

```
## 781128            77560                  77560             0
## 781129            85260                  85260             0
## 781130            83440                  83440             0
## 781131            77560                  77560             0
## 781132            77560                  77560             0
##        appraised_building appraised_outbuilding appraised_extra_feature
## 781127             110800                     0                      NA
## 781128             110800                     0                      NA
## 781129             121800                     0                      NA
## 781130             119200                     0                      NA
## 781131             110800                     0                      NA
## 781132             110800                     0                      NA
##        valuation_year zone zone_description model condition
## 781127           2021  RM1             <NA>     5         A
## 781128           2021  RM1             <NA>     5         A
## 781129           2021  RM1             <NA>     5         A
## 781130           2021  RM1             <NA>     5         A
## 781131           2021  RM1             <NA>     5         A
## 781132           2021  RM1             <NA>     5         A
##        condition_description  ayb  eyb living_area effective_area total_rooms
## 781127               Average 1986 2005         977            977           4
## 781128               Average 1986 2005         977            977           4
## 781129               Average 1986 2005        1166           1166           4
## 781130               Average 1986 2005        1130           1130           4
## 781131               Average 1985 2005         977            977           4
## 781132               Average 1986 2005         977            977           4
##        number_of_bedroom number_of_baths number_of_half_baths occupancy
## 781127                 2               1                    1         1
## 781128                 2               1                    1         1
## 781129                 2               1                    0         1
## 781130                 2               1                    0         1
## 781131                 2               1                    1         1
## 781132                 2               1                    1         1
##        sale_price           sale_date qualified     prior_sale_date
## 781127     117000 1987-08-20 20:00:00         U               <NA>
## 781128      96500 2018-06-07 20:00:00         Q 2007-05-24 20:00:00
## 781129      63000 1994-10-02 20:00:00         Q 1986-04-27 20:00:00
## 781130          0 1993-05-25 20:00:00         U               <NA>
## 781131          0 2018-02-21 19:00:00         U 1986-04-28 20:00:00
## 781132          0 2023-04-19 20:00:00         U 1998-05-28 20:00:00
##        prior_book_page prior_sale_price editor edit_date collection_year
## 781127            <NA>               NA   <NA>      <NA>            2023
## 781128       7965/0056           165000   <NA>      <NA>            2023
## 781129       3445/0113                0   <NA>      <NA>            2023
## 781130            <NA>               NA   <NA>      <NA>            2023
## 781131       3445/0248                0   <NA>      <NA>            2023
## 781132       5317/0333            51000   <NA>      <NA>            2023
##        planning_region state_use state_use_description
## 781127   South Central      1020           Condominium
## 781128   South Central      1020           Condominium
## 781129   South Central      1020           Condominium
## 781130   South Central      1020           Condominium
## 781131   South Central      1020           Condominium
## 781132   South Central      1020           Condominium
```

```
##                                       globalid shape_length shape_area
## 781127 {0D9FFBDB-339A-4586-99BC-DF75B78F086F}     483.8936   12509.72
## 781128 {6EB553F3-9ECD-475F-9358-DA65DB1D90B4}     483.8936   12509.72
## 781129 {BF991809-1D3F-491D-BB28-5FD7EA57454B}     483.8936   12509.72
## 781130 {3DB64A53-C31B-41A6-8009-AC58703A47F5}     483.8936   12509.72
## 781131 {F65B7056-4447-4DA2-A762-EFB2E3CF3DE6}     483.8936   12509.72
## 781132 {3B70C5E0-4DC3-403C-A010-06E69939A886}     483.8936   12509.72
```

```r
# Create `sf` variable indicating single-family if description matches
d$sf <- ifelse(tolower(d$state_use_description) %in%
                 c(single_desc, apt_desc, condo_desc), 1, 0)

# Verify if the classification works correctly
table(d$sf, d$town_name, useNA = 'always')
```

```
##
##        NEW HAVEN  <NA>
##   0         4855     0
##   1        11809     0
##   <NA>         0     0
```

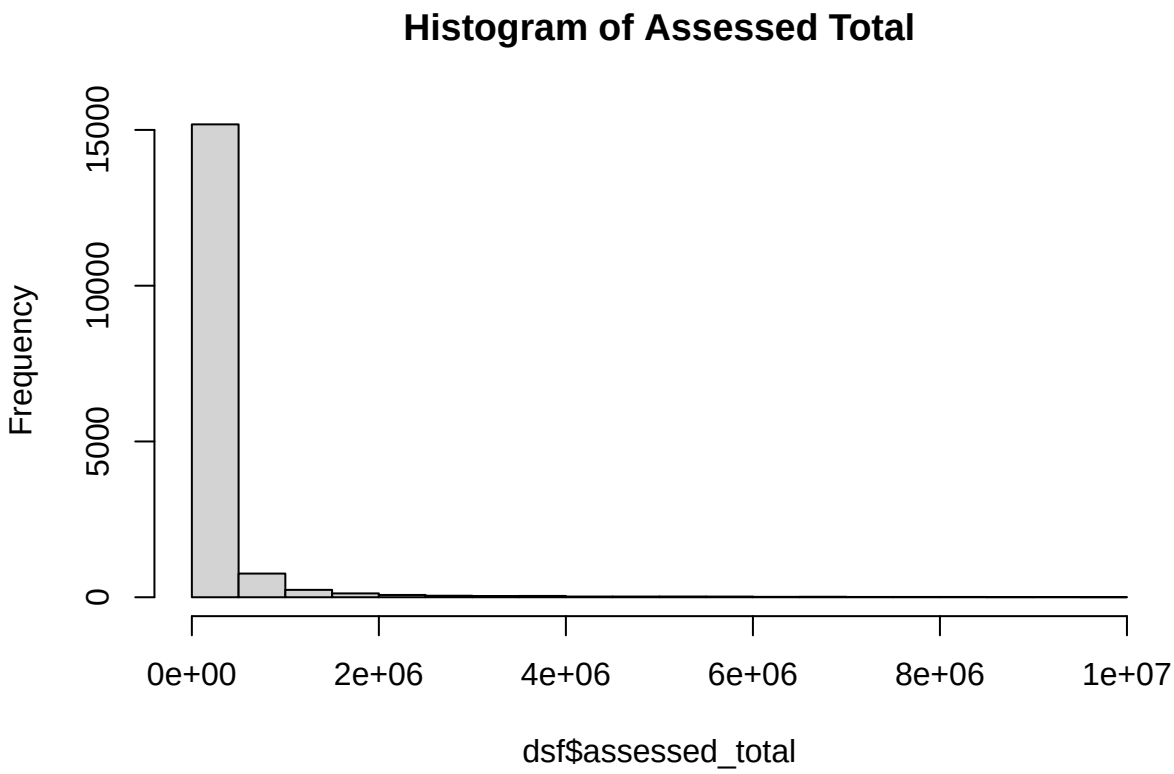
We conclude that `Residential` and `Condominiums` should be considered as single-family homes. This decision aligns with earlier restrictions applied to limit the number of rooms, bathrooms, and assessed total.

## Step (3): Using EDA to further clean data

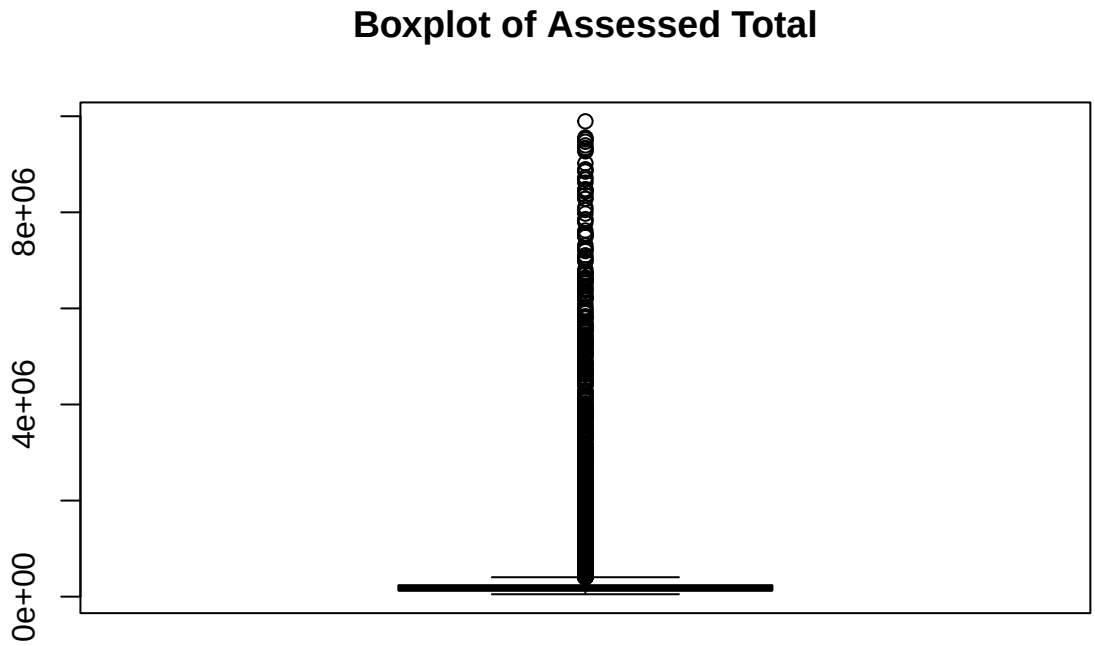
### Assessed Total and Log-Transformation

Applying a log transformation on the `assessed_total` variable helps normalize the data distribution and enhances model performance.

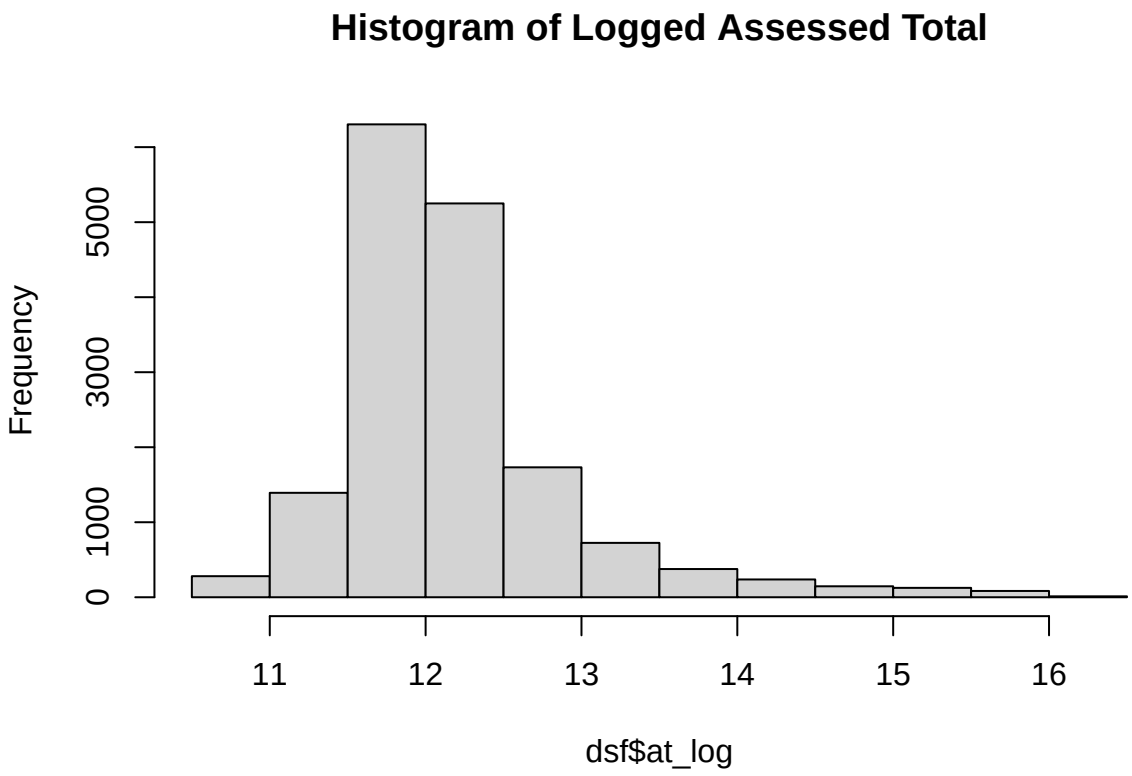
```r
# Save a new dataframe to isolate issues during the cleaning stage
dsf <- d

# Visualize the distribution of `assessed_total` before transformation
hist(dsf$assessed_total, main = "Histogram of Assessed Total")
```




```
boxplot(dsf$assessed_total, main = "Boxplot of Assessed Total")
```




```
# Apply log transformation
dsf$at_log <- log(dsf$assessed_total)
hist(dsf$at_log, main = "Histogram of Logged Assessed Total")
```

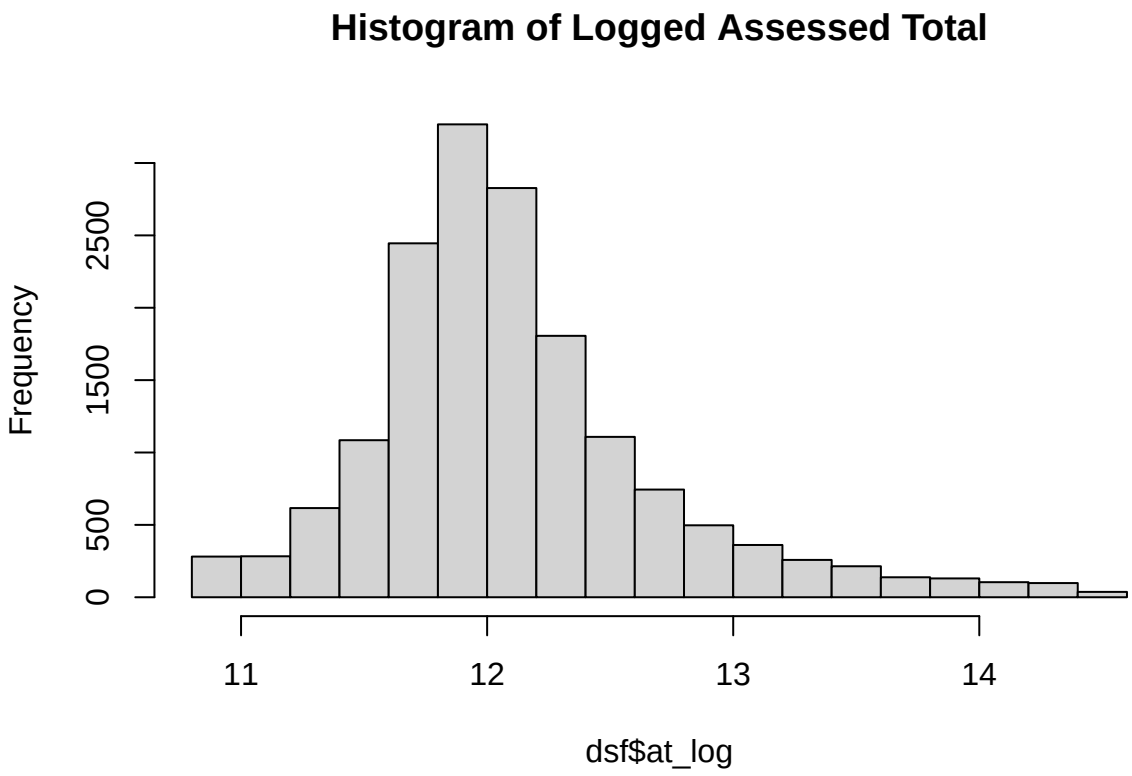



Observing the distribution, a long tail remains after transformation. We filter out properties priced over $2 million to remove extreme values, which are not suitable for single family houses.

```
# Filter properties priced above $2 million
sum(dsf$assessed_total > 2000000)
```

```
## [1] 365
```

```
dsf <- dsf[dsf$assessed_total < 2000000, ]

# Visualize the improved distribution after filtering
hist(dsf$at_log, main = "Histogram of Logged Assessed Total")
```




### Number of Rooms

Analyze room-related variables like bedrooms, bathrooms, and total rooms.

```
# Verify if any properties have 0 total rooms (should not be the case)
nrow(dsf)
```

```
## [1] 16299
```

```
sum(dsf$total_rooms == 0, na.rm = T)
```

```
## [1] 13
```

```
range(dsf$total_rooms, na.rm = T)
```

```
## [1]  0 82
```

```
# checked  480 WOODWARD AV : does not look like a home
# checked 83 EAST GRAND AV : it is a 4 bed - 2 bath home
# checked 70 FOXON ST: it is a container
# checked 41 TRUMBULL ST: it is a  2 bed - 1 bath apartment
# it might be possible that depending on the floor of this property,
# it may indicate basement and hence has no room
```

```r
# checked 842 HOWARD AV through vision solutions:
# we believe that it is using an algorithm to fill information
# we do not believe that this is accurate
head(dsf[dsf$total_rooms == 0 & !is.na(dsf$total_rooms), ])
```

```
##        town_name                link                       owner co_owner
## 755982 NEW HAVEN 52070-053 0950 00800    EASTROCK INVESTMENTS LLC     <NA>
## 758417 NEW HAVEN 52070-098 1008 01800              G&B REALTY LLC     <NA>
## 759141 NEW HAVEN 52070-118 1035 00900              DELUCIA ANGELO     <NA>
## 764997 NEW HAVEN 52070-223 0380 00500  PM TRUMBULL ASSOCIATES LLC     <NA>
## 771033 NEW HAVEN 52070-278 0149 01000 YALE-NEW HAVEN HOSPITAL INC     <NA>
## 772546 NEW HAVEN 52070-286 0433 01501            DIBENEDETTO KIM      <NA>
##                location  mailing_address mailing_city mailing_state
## 755982 480 WOODWARD AV  480 WOODWARD AV     NEW HAVEN            CT
## 758417 83 EAST GRAND AV 17 CARRINGTON DR    GREENWICH            CT
## 759141     70 FOXON ST     130 WARNER RD   EAST HAVEN            CT
## 764997   41 TRUMBULL ST   41 TRUMBULL ST    NEW HAVEN            CT
## 771033    842 HOWARD AV       20 YORK ST    NEW HAVEN            CT
## 772546    89 NEWHALL ST    89 NEWHALL ST    NEW HAVEN            CT
##        assessed_total assessed_land assessed_building pre_year_assessed_total
## 755982        1004010        245000            723310                 1004010
## 758417         936600        227500            696360                  936600
## 759141         161700         63770             91980                  161700
## 764997         474040        174580            296940                  474040
## 771033         267680        122710            144970                  267680
## 772546         181160         34440            146720                  181160
##        appraised_land appraised_building appraised_outbuilding
## 755982         350000            1033300                 51000
## 758417         325000             994800                 18200
## 759141          91100             131400                  8500
## 764997         249400             424200                     0
## 771033         175300             207100                     0
## 772546          49200             209600                     0
##        appraised_extra_feature valuation_year   zone zone_description model
## 755982                      NA           2021    RM1             <NA>    94
## 758417                      NA           2021    RM1             <NA>    94
## 759141                      NA           2021     BA             <NA>    95
## 764997                      NA           2021     RO             <NA>    94
## 771033                      NA           2021     RO             <NA>    94
## 772546                      NA           2021 PDD 49             <NA>     1
##        condition condition_description  ayb  eyb living_area effective_area
## 755982         A               Average 1966 1981       13104          14448
## 758417        VG             Very Good 1979 1999        6386           6451
## 759141         A               Average 2003 2011        2802           2802
## 764997         A               Average 1900 1981        5190           5225
## 771033         A               Average 1890 1981        2430           2836
## 772546         G                  Good 1998 2012        1568           1889
##        total_rooms number_of_bedroom number_of_baths number_of_half_baths
## 755982           0                NA              NA                   NA
## 758417           0                NA              NA                   NA
## 759141           0                NA              NA                   NA
## 764997           0                NA              NA                   NA
## 771033           0                NA              NA                   NA
```

```
## 772546          0                3              2                   1
##        occupancy sale_price           sale_date qualified     prior_sale_date
## 755982        14          0 2021-08-09 20:00:00         U 2002-12-12 19:00:00
## 758417         9     595000 2013-05-05 20:00:00         Q 2010-08-05 20:00:00
## 759141         1      54000 2001-10-28 19:00:00         U                <NA>
## 764997         6     300000 2001-02-21 19:00:00         Q 1964-06-21 20:00:00
## 771033         2          0 2012-09-04 20:00:00         U 1990-05-31 20:00:00
## 772546         1      74900 1999-07-22 20:00:00      <NA>                <NA>
##        prior_book_page prior_sale_price editor edit_date collection_year
## 755982       6280/0051           575000   <NA>      <NA>            2023
## 758417       8582/0195                0   <NA>      <NA>            2023
## 759141             0/0                0   <NA>      <NA>            2023
## 764997             0/0                0   <NA>      <NA>            2023
## 771033       4254/0015           400000   <NA>      <NA>            2023
## 772546            <NA>               NA   <NA>      <NA>            2023
##        planning_region state_use state_use_description
## 755982   South Central      <NA>    APT Over12  MDL-94
## 758417   South Central      <NA>    APT Over12  MDL-94
## 759141   South Central      3320    SVC SHP/GA  MDL-95
## 764997   South Central      3400    OFFICE BLD  MDL-94
## 771033   South Central      <NA>      PVT HOSP  MDL-96
## 772546   South Central      1010         Single Family
##                                    globalid shape_length shape_area sf
## 755982 {7A05BEF4-8F0B-49D1-9A8C-4BD36D0C8F9F}     410.9682  7962.1614  1
## 758417 {800F3840-59EF-4E82-9E38-70F543071859}     314.7243  5580.8030  1
## 759141 {93200A26-0CE3-468F-A689-5656DD14C95E}     149.9778  1306.0684  0
## 764997 {46FED1AD-770B-4552-8B8E-5080D452322C}     108.9173   562.2555  0
## 771033 {C26F7CA1-D0F2-4E25-8CF9-C3DFABE201B7}     130.3909   713.6486  0
## 772546 {28AF8FCB-D981-4A59-8283-07F3D217304C}     138.3014  1165.7288  1
##          at_log
## 755982 13.81951
## 758417 13.75001
## 759141 11.99350
## 764997 13.06905
## 771033 12.49755
## 772546 12.10714
```

```
# Remove entries with 0 total rooms due to data issues or non-residential use
dsf <- dsf[!(dsf$total_rooms == 0) | is.na(dsf$total_rooms), ]

# Remove properties with more than 10 total rooms as they may be outliers
sum(dsf$total_rooms > 10, na.rm = T)
```

```
## [1] 594
```

```
dsf <- dsf[!(dsf$total_rooms > 10) | is.na(dsf$total_rooms), ]

# Verify the range after filtering
range(dsf$total_rooms, na.rm = T)
```

```
## [1]  1 10
```

```r
# Check for NA values in room-related variables
sum(is.na(dsf$total_rooms))
```

```
## [1] 2112
```

```r
sum(is.na(dsf$number_of_bedroom))
```

```
## [1] 2017
```

```r
sum(is.na(dsf$number_of_baths))
```

```
## [1] 2017
```

```r
sum(is.na(dsf$number_of_half_baths))
```

```
## [1] 2054
```

```r
# Check if NA values are in one or multiple columns for consistency
# That is, can we fix issues from other columns if we fix one?
# This may indicate that many of the missing value problmes may be an
# artifact of bad properties
temp <- dsf[!is.na(dsf$number_of_baths), ]
sum(is.na(temp$number_of_bedroom))
```

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

```r
sum(is.na(temp$number_of_half_baths))
```

```
## [1] 38
```

```r
# Review entries with missing bedroom/bathroom values
head(dsf[is.na(dsf$number_of_bedroom), ])
```

```
##        town_name                 link                              owner
## 754308 NEW HAVEN 52070-015 0876 00100                        VPS REALTY LLC
## 754331 NEW HAVEN 52070-019 0900 00100                     CITY OF NEW HAVEN
## 754889 NEW HAVEN 52070-027 0904 00800 CITY OF NEW HAVEN HOUSING AUTHORITY
## 755030 NEW HAVEN 52070-028 0898 00100       DELMONACO ASSUNTA & TERIGIO &
## 755147 NEW HAVEN 52070-029 0893 00500                     65 BURR STREET LLC
## 755383 NEW HAVEN 52070-032 0872 01000 CITY OF NEW HAVEN HOUSING AUTHORITY
##                 co_owner          location      mailing_address mailing_city
## 754308              <NA>    25 TOWNSEND AV 984 ROUTE 9 SUITE A9       PARLIN
## 754331              <NA>      191 BURR ST         165 CHURCH ST    NEW HAVEN
## 754889 CITY OF NEW HAVEN 121 STUYVESANT AV         P O BOX 1912    NEW HAVEN
## 755030              SURV   473 TOWNSEND AV      473 TOWNSEND AV    NEW HAVEN
## 755147              <NA>        65 BURR ST         15 SANDRA DR     BRANFORD
## 755383 CITY OF NEW HAVEN     6 TOWNSEND AV         P O BOX 1917    NEW HAVEN
##        mailing_state assessed_total assessed_land assessed_building
## 754308            NJ         308210         81620            205310
```

```
## 754331            CT        1067990             0           1029630
## 754889            CT         255150         64680            184030
## 755030            CT         276500         70210            203070
## 755147            CT         180110         63910            109760
## 755383            CT         181090         63840            117250
##        pre_year_assessed_total appraised_land appraised_building
## 754308                  308210         116600             293300
## 754331                  899500              0            1470900
## 754889                  255150          92400             262900
## 755030                  276500         100300             290100
## 755147                  180110          91300             156800
## 755383                  181090          91200             167500
##        appraised_outbuilding appraised_extra_feature valuation_year    zone
## 754308                 30400                      NA           2021      BA
## 754331                     0                      NA           2021 AIRPORT
## 754889                  9200                      NA           2021     RS2
## 755030                     0                      NA           2021     RS2
## 755147                  9200                      NA           2021     RS2
## 755383                     0                      NA           2021     RS2
##        zone_description model condition condition_description  ayb  eyb
## 754308             <NA>    94         A               Average 1976 1986
## 754331             <NA>    94         G                  Good 1930 1986
## 754889             <NA>    94         G                  Good 1956 1986
## 755030             <NA>    94         A               Average 1926 1981
## 755147             <NA>    94         F                     F 1900 1976
## 755383             <NA>    94         A               Average 1947 1981
##        living_area effective_area total_rooms number_of_bedroom number_of_baths
## 754308        2784           2921          NA                NA              NA
## 754331       12236          13259          NA                NA              NA
## 754889        2095           2288          NA                NA              NA
## 755030        4140           4478          NA                NA              NA
## 755147        2736           2862          NA                NA              NA
## 755383        1326           1552          NA                NA              NA
##        number_of_half_baths occupancy sale_price           sale_date qualified
## 754308                   NA         1     370000 2001-11-28 19:00:00         U
## 754331                   NA         1      85000 1995-02-21 19:00:00         U
## 754889                   NA         1     137800 1993-01-10 19:00:00      <NA>
## 755030                   NA         2     120000 1992-03-23 19:00:00      <NA>
## 755147                   NA         2     204800 2018-02-11 19:00:00         Q
## 755383                   NA         1     115500 1993-01-13 19:00:00         Q
##            prior_sale_date prior_book_page prior_sale_price editor edit_date
## 754308 1986-06-03 20:00:00       3468/0240           180000   <NA>      <NA>
## 754331               <NA>            <NA>               NA   <NA>      <NA>
## 754889               <NA>            <NA>               NA   <NA>      <NA>
## 755030 1987-08-10 20:00:00       3732/0250           240000   <NA>      <NA>
## 755147 2011-09-22 20:00:00       8734/0209           185000   <NA>      <NA>
## 755383               <NA>            <NA>               NA   <NA>      <NA>
##        collection_year planning_region state_use state_use_description
## 754308            2023   South Central      3220    STORE/SHOP  MDL-94
## 754331            2023   South Central      9020          CITY  MDL-94
## 754889            2023   South Central      9080     HSNG AUTH  MDL-94
## 755030            2023   South Central      3030     MIXED USE  MDL-94
## 755147            2023   South Central      3030     MIXED USE  MDL-94
## 755383            2023   South Central      9080     HSNG AUTH  MDL-94
```

```
##                                      globalid shape_length shape_area sf
## 754308 {74C94981-9D1F-4237-92E2-29BCA1EA7833}     225.6785   3206.886  0
## 754331 {147BBAA6-4B6A-4248-B623-1E96D68B76A8}    1055.6852  62010.509  0
## 754889 {BA2B6961-D3BA-462A-AA01-B9CDDBDE8809}     182.3970   2058.857  0
## 755030 {39711480-90DE-4733-96D0-E90A97A5D7DA}     206.3201   2560.910  0
## 755147 {FB61E59C-C9EE-4CAF-BD09-AE83E0D1D1E4}     137.4196   1144.618  0
## 755383 {D7EF5673-C424-42A4-8595-A198109B347D}     146.3244   1084.511  0
##          at_log
## 754308 12.63854
## 754331 13.88129
## 754889 12.44961
## 755030 12.52997
## 755147 12.10132
## 755383 12.10675
```

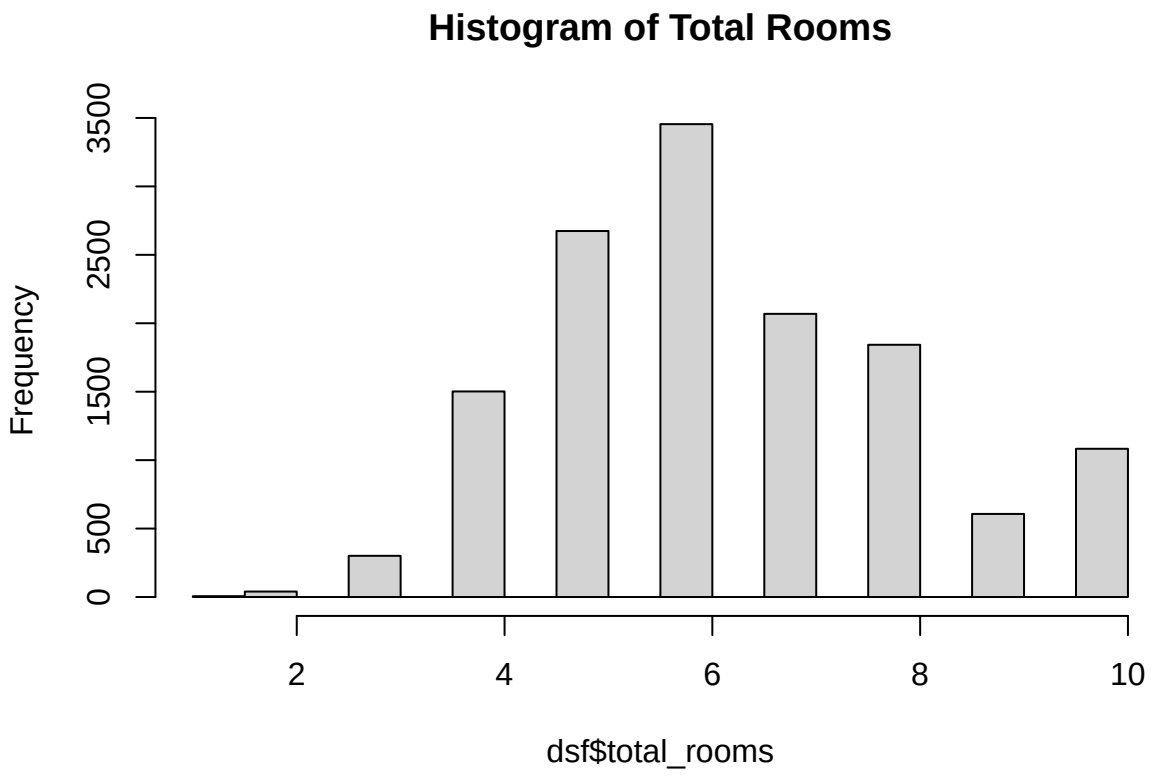
```
head(dsf[is.na(dsf$number_of_baths), ])
```

```
##        town_name                 link                              owner
## 754308 NEW HAVEN 52070-015 0876 00100                       VPS REALTY LLC
## 754331 NEW HAVEN 52070-019 0900 00100                    CITY OF NEW HAVEN
## 754889 NEW HAVEN 52070-027 0904 00800 CITY OF NEW HAVEN HOUSING AUTHORITY
## 755030 NEW HAVEN 52070-028 0898 00100       DELMONACO ASSUNTA & TERIGIO &
## 755147 NEW HAVEN 52070-029 0893 00500                    65 BURR STREET LLC
## 755383 NEW HAVEN 52070-032 0872 01000 CITY OF NEW HAVEN HOUSING AUTHORITY
##                 co_owner          location      mailing_address mailing_city
## 754308              <NA>    25 TOWNSEND AV 984 ROUTE 9 SUITE A9       PARLIN
## 754331              <NA>      191 BURR ST         165 CHURCH ST    NEW HAVEN
## 754889 CITY OF NEW HAVEN 121 STUYVESANT AV         P O BOX 1912    NEW HAVEN
## 755030              SURV   473 TOWNSEND AV      473 TOWNSEND AV    NEW HAVEN
## 755147              <NA>        65 BURR ST         15 SANDRA DR     BRANFORD
## 755383 CITY OF NEW HAVEN     6 TOWNSEND AV         P O BOX 1917    NEW HAVEN
##        mailing_state assessed_total assessed_land assessed_building
## 754308            NJ         308210         81620            205310
## 754331            CT        1067990             0           1029630
## 754889            CT         255150         64680            184030
## 755030            CT         276500         70210            203070
## 755147            CT         180110         63910            109760
## 755383            CT         181090         63840            117250
##        pre_year_assessed_total appraised_land appraised_building
## 754308                  308210         116600             293300
## 754331                  899500              0            1470900
## 754889                  255150          92400             262900
## 755030                  276500         100300             290100
## 755147                  180110          91300             156800
## 755383                  181090          91200             167500
##        appraised_outbuilding appraised_extra_feature valuation_year    zone
## 754308                 30400                      NA           2021      BA
## 754331                     0                      NA           2021 AIRPORT
## 754889                  9200                      NA           2021     RS2
## 755030                     0                      NA           2021     RS2
## 755147                  9200                      NA           2021     RS2
## 755383                     0                      NA           2021     RS2
##        zone_description model condition condition_description  ayb  eyb
## 754308             <NA>    94         A               Average 1976 1986
```

```
## 754331             <NA>    94         G                 Good 1930 1986
## 754889             <NA>    94         G                 Good 1956 1986
## 755030             <NA>    94         A              Average 1926 1981
## 755147             <NA>    94         F                    F 1900 1976
## 755383             <NA>    94         A              Average 1947 1981
##        living_area effective_area total_rooms number_of_bedroom number_of_baths
## 754308        2784           2921          NA                NA              NA
## 754331       12236          13259          NA                NA              NA
## 754889        2095           2288          NA                NA              NA
## 755030        4140           4478          NA                NA              NA
## 755147        2736           2862          NA                NA              NA
## 755383        1326           1552          NA                NA              NA
##        number_of_half_baths occupancy sale_price           sale_date qualified
## 754308                   NA         1     370000 2001-11-28 19:00:00         U
## 754331                   NA         1      85000 1995-02-21 19:00:00         U
## 754889                   NA         1     137800 1993-01-10 19:00:00      <NA>
## 755030                   NA         2     120000 1992-03-23 19:00:00      <NA>
## 755147                   NA         2     204800 2018-02-11 19:00:00         Q
## 755383                   NA         1     115500 1993-01-13 19:00:00         Q
##           prior_sale_date prior_book_page prior_sale_price editor edit_date
## 754308 1986-06-03 20:00:00       3468/0240           180000   <NA>      <NA>
## 754331                <NA>            <NA>               NA   <NA>      <NA>
## 754889                <NA>            <NA>               NA   <NA>      <NA>
## 755030 1987-08-10 20:00:00       3732/0250           240000   <NA>      <NA>
## 755147 2011-09-22 20:00:00       8734/0209           185000   <NA>      <NA>
## 755383                <NA>            <NA>               NA   <NA>      <NA>
##        collection_year planning_region state_use state_use_description
## 754308            2023   South Central      3220    STORE/SHOP  MDL-94
## 754331            2023   South Central      9020          CITY  MDL-94
## 754889            2023   South Central      9080     HSNG AUTH  MDL-94
## 755030            2023   South Central      3030     MIXED USE  MDL-94
## 755147            2023   South Central      3030     MIXED USE  MDL-94
## 755383            2023   South Central      9080     HSNG AUTH  MDL-94
##                                    globalid shape_length shape_area sf
## 754308 {74C94981-9D1F-4237-92E2-29BCA1EA7833}     225.6785   3206.886  0
## 754331 {147BBAA6-4B6A-4248-B623-1E96D68B76A8}    1055.6852  62010.509  0
## 754889 {BA2B6961-D3BA-462A-AA01-B9CDDBDE8809}     182.3970   2058.857  0
## 755030 {39711480-90DE-4733-96D0-E90A97A5D7DA}     206.3201   2560.910  0
## 755147 {FB61E59C-C9EE-4CAF-BD09-AE83E0D1D1E4}     137.4196   1144.618  0
## 755383 {D7EF5673-C424-42A4-8595-A198109B347D}     146.3244   1084.511  0
##          at_log
## 754308 12.63854
## 754331 13.88129
## 754889 12.44961
## 755030 12.52997
## 755147 12.10132
## 755383 12.10675
```

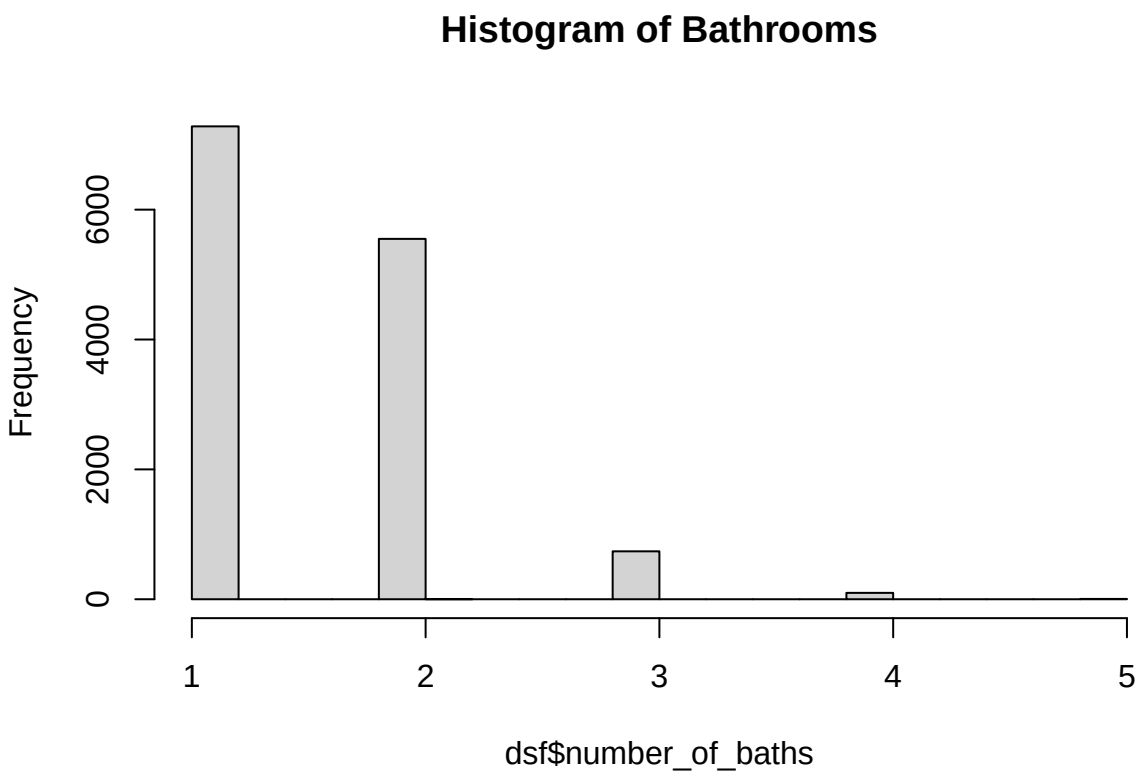
Visualize room variable distributions

```r
hist(dsf$total_rooms, main = "Histogram of Total Rooms")
```




```r
hist(dsf$number_of_bedroom, main = "Histogram of Bedrooms")
```


```r
hist(dsf$number_of_baths, main = "Histogram of Bathrooms")
```




Verify if `total_rooms` equals the sum of other room variables: it doesn't

```
head(dsf[, c('total_rooms', 'number_of_bedroom',
             'number_of_baths', 'number_of_half_baths')], 5)
```

```
##        total_rooms number_of_bedroom number_of_baths number_of_half_baths
## 754265           6                 3               2                    1
## 754266           6                 3               2                    1
## 754267           8                 3               1                    0
## 754268           5                 3               1                    0
## 754270           6                 3               1                    1
```

**Property Size: Living Area and Effective Area**

```
# Check for missing values in `living_area` and `effective_area`
nrow(dsf[is.na(dsf$living_area), ])  # Missing values in living_area
```

```
## [1] 27
```

```
nrow(dsf[is.na(dsf$effective_area), ])  # Missing values in effective_area
```

```
## [1] 27
```

```r
# In some cases, living area matches effective area. To maintain consistency,
# focus on `living_area` and disregard `effective_area` to reduce NAs.
head(dsf[, c("living_area", "effective_area")])
```

```
##        living_area effective_area
## 754265        1475           1706
## 754266        1792           1996
## 754267         864           1129
## 754268        1040           1289
## 754270        1080           1238
## 754271        1040           1248
```

```r
tail(dsf[, c("living_area", "effective_area")])
```

```
##        living_area effective_area
## 781272        1752           2020
## 781276          NA             NA
## 781284        4100           4264
## 781286       22000          22000
## 781287        6035           6786
## 781296        2960           2960
```

```r
# Living Area: Check the range and remove entries outside reasonable thresholds
range(dsf$living_area, na.rm = T)
```

```
## [1]    231 336400
```

```r
sum(dsf$living_area < 200, na.rm = T)
```

```
## [1] 0
```

```r
sum(dsf$living_area > 10000, na.rm = T)
```

```
## [1] 395
```

```r
# Effective Area: Check the range and remove entries outside reasonable thresholds
range(dsf$effective_area, na.rm = T)
```

```
## [1]    231 336400
```

```r
sum(dsf$effective_area < 200, na.rm = T)
```

```
## [1] 0
```

```r
sum(dsf$effective_area > 10000, na.rm = T)
```

```
## [1] 432
```

```r
# Checked 191 BURR S: a moderately big building and a huge lawn
# Checked 560 WOODWARD AV: a moderately big building and a huge lawn
# Checked 821 EAST SHORE PKWY: this is a huge barren land with buildings
head(dsf[dsf$living_area > 10000 & !is.na(dsf$living_area), ])
```

```
##        town_name                link                          owner
## 754331 NEW HAVEN 52070-019 0900 00100             CITY OF NEW HAVEN
## 756018 NEW HAVEN 52070-053 0950 01900  ANNEX YOUNG MENS ASSOCIATION T
## 756488 NEW HAVEN 52070-060 0947 01500 SOUTHERN NEW ENGLAND TELEPHONE
## 756731 NEW HAVEN 52070-067 0953 00300   15 STILES STREET CORPORATION
## 756732 NEW HAVEN 52070-067 0953 00400               KOOPER GROUP LLC
## 756756 NEW HAVEN 52070-069 0955 00300                 BS KENDALL LLC
##                          co_owner           location
## 754331                       <NA>        191 BURR ST
## 756018          C/O ANTHONY RUOCCO    560 WOODWARD AV
## 756488 C/O FRONTIER COMMUNICATIONS    1155 TOWNSEND AV
## 756731                       <NA> 821 EAST SHORE PKWY
## 756732                       <NA>       50 FULTON TER
## 756756                       <NA>      111 KENDALL ST
##                     mailing_address mailing_city mailing_state assessed_total
## 754331                165 CHURCH ST    NEW HAVEN            CT        1067990
## 756018                   PO BOX 157   EAST HAVEN            CT         731430
## 756488                  PO BOX 2629      ADDISON            TX         861840
## 756731            C/O MICHAEL WEINER   WOODBRIDGE            CT         697060
## 756732            C/O MICHAEL WEINER   WOODBRIDGE            CT         436660
## 756756 C/O STEVE SADLER & BILL THOMAS    NEW HAVEN            CT         460320
##        assessed_land assessed_building pre_year_assessed_total appraised_land
## 754331             0           1029630                  899500             0
## 756018        161700            512330                  731430        231000
## 756488         53550            808290                  861840         76500
## 756731        170100            291410                  697060        243000
## 756732        143010            215460                  436660        204300
## 756756         38220            398510                  460320         54600
##        appraised_building appraised_outbuilding appraised_extra_feature
## 754331            1470900                    0                      NA
## 756018             731900                82000                      NA
## 756488            1154700                    0                      NA
## 756731             416300                26500                      NA
## 756732             307800               100000                      NA
## 756756             569300                 4700                      NA
##        valuation_year    zone zone_description model condition
## 754331           2021 AIRPORT            <NA>    94         G
## 756018           2021     RM1            <NA>    94         A
## 756488           2021     RM1            <NA>    96         A
## 756731           2021      IH            <NA>    96         F
## 756732           2021      IH            <NA>    96         A
## 756756           2021      IH            <NA>    96         G
##        condition_description  ayb  eyb living_area effective_area total_rooms
## 754331                  Good 1930 1986       12236          13259          NA
## 756018               Average 1950 1981       14052          14052          NA
## 756488               Average 1925 1981       17872          17872          NA
## 756731                     F 1960 1976       10600          10600          NA
## 756732               Average 1952 1981       11215          11496          NA
```

```
## 756756                  Good 1963 1986       10380          12990          NA
##        number_of_bedroom number_of_baths number_of_half_baths occupancy
## 754331                NA              NA                   NA         1
## 756018                NA              NA                   NA         1
## 756488                NA              NA                   NA         1
## 756731                NA              NA                   NA         1
## 756732                NA              NA                   NA         1
## 756756                NA              NA                   NA         1
##        sale_price           sale_date qualified     prior_sale_date
## 754331      85000 1995-02-21 19:00:00         U                <NA>
## 756018          0                <NA>      <NA>                <NA>
## 756488          0 1899-12-31 19:00:00         U                <NA>
## 756731     320160 1995-05-08 20:00:00         Q 1992-07-19 20:00:00
## 756732     787929 2009-03-17 20:00:00         Q 1995-05-08 20:00:00
## 756756     450000 2015-12-28 19:00:00         Q 2010-09-07 20:00:00
##        prior_book_page prior_sale_price editor edit_date collection_year
## 754331            <NA>               NA   <NA>      <NA>            2023
## 756018            <NA>               NA   <NA>      <NA>            2023
## 756488            <NA>               NA   <NA>      <NA>            2023
## 756731       4505/0324                0   <NA>      <NA>            2023
## 756732       4861/0269            83160   <NA>      <NA>            2023
## 756756       8595/0319           450000   <NA>      <NA>            2023
##        planning_region state_use state_use_description
## 754331   South Central      9020          CITY  MDL-94
## 756018   South Central      3530    FRATNL ORG  MDL-94
## 756488   South Central      4300     TEL X STA  MDL-96
## 756731   South Central      4010     IND WHSES  MDL-96
## 756732   South Central      4010     IND WHSES  MDL-96
## 756756   South Central      4000    INDUSTRIAL  MDL-96
##                                      globalid shape_length shape_area sf
## 754331 {147BBAA6-4B6A-4248-B623-1E96D68B76A8}    1055.6852  62010.509  0
## 756018 {7218298A-C821-4D28-8471-369CAE60465D}     632.6066  16165.154  0
## 756488 {EFD5056D-6D04-482F-B5FE-62BB41AB8EF4}     227.9482   2817.456  0
## 756731 {E8034B45-CCFE-4D34-A78F-CB54FBF60624}     463.8206   5836.079  0
## 756732 {53C79594-04D9-419A-B245-36C5BCEC2637}     232.8151   3311.196  0
## 756756 {6B78162F-05FE-4CDE-8474-E0754C3101C6}     194.0787   2292.230  0
##          at_log
## 754331 13.88129
## 756018 13.50276
## 756488 13.66682
## 756731 13.45463
## 756732 12.98691
## 756756 13.03968
```

```
# Using effective area return similar properties
head(dsf[dsf$effective_area > 10000 & !is.na(dsf$effective_area), ])
```

```
##        town_name                link                          owner
## 754331 NEW HAVEN 52070-019 0900 00100             CITY OF NEW HAVEN
## 756018 NEW HAVEN 52070-053 0950 01900  ANNEX YOUNG MENS ASSOCIATION T
## 756488 NEW HAVEN 52070-060 0947 01500 SOUTHERN NEW ENGLAND TELEPHONE
## 756731 NEW HAVEN 52070-067 0953 00300   15 STILES STREET CORPORATION
## 756732 NEW HAVEN 52070-067 0953 00400               KOOPER GROUP LLC
## 756756 NEW HAVEN 52070-069 0955 00300                 BS KENDALL LLC
```

```
##                           co_owner            location
## 754331                        <NA>        191 BURR ST
## 756018          C/O ANTHONY RUOCCO    560 WOODWARD AV
## 756488 C/O FRONTIER COMMUNICATIONS    1155 TOWNSEND AV
## 756731                        <NA> 821 EAST SHORE PKWY
## 756732                        <NA>       50 FULTON TER
## 756756                        <NA>      111 KENDALL ST
##                      mailing_address mailing_city mailing_state assessed_total
## 754331                 165 CHURCH ST    NEW HAVEN            CT        1067990
## 756018                    PO BOX 157   EAST HAVEN            CT         731430
## 756488                  PO BOX 2629      ADDISON            TX         861840
## 756731            C/O MICHAEL WEINER   WOODBRIDGE            CT         697060
## 756732            C/O MICHAEL WEINER   WOODBRIDGE            CT         436660
## 756756 C/O STEVE SADLER & BILL THOMAS    NEW HAVEN            CT         460320
##        assessed_land assessed_building pre_year_assessed_total appraised_land
## 754331             0           1029630                  899500              0
## 756018        161700            512330                  731430         231000
## 756488         53550            808290                  861840          76500
## 756731        170100            291410                  697060         243000
## 756732        143010            215460                  436660         204300
## 756756         38220            398510                  460320          54600
##        appraised_building appraised_outbuilding appraised_extra_feature
## 754331            1470900                    0                      NA
## 756018             731900                82000                      NA
## 756488            1154700                    0                      NA
## 756731             416300                26500                      NA
## 756732             307800               100000                      NA
## 756756             569300                 4700                      NA
##        valuation_year    zone zone_description model condition
## 754331           2021 AIRPORT             <NA>    94         G
## 756018           2021     RM1             <NA>    94         A
## 756488           2021     RM1             <NA>    96         A
## 756731           2021      IH             <NA>    96         F
## 756732           2021      IH             <NA>    96         A
## 756756           2021      IH             <NA>    96         G
##        condition_description  ayb  eyb living_area effective_area total_rooms
## 754331                  Good 1930 1986       12236          13259          NA
## 756018               Average 1950 1981       14052          14052          NA
## 756488               Average 1925 1981       17872          17872          NA
## 756731                     F 1960 1976       10600          10600          NA
## 756732               Average 1952 1981       11215          11496          NA
## 756756                  Good 1963 1986       10380          12990          NA
##        number_of_bedroom number_of_baths number_of_half_baths occupancy
## 754331                NA              NA                   NA         1
## 756018                NA              NA                   NA         1
## 756488                NA              NA                   NA         1
## 756731                NA              NA                   NA         1
## 756732                NA              NA                   NA         1
## 756756                NA              NA                   NA         1
##        sale_price           sale_date qualified     prior_sale_date
## 754331      85000 1995-02-21 19:00:00         U               <NA>
## 756018          0               <NA>      <NA>               <NA>
## 756488          0 1899-12-31 19:00:00         U               <NA>
## 756731     320160 1995-05-08 20:00:00         Q 1992-07-19 20:00:00
```

```
## 756732     787929 2009-03-17 20:00:00         Q 1995-05-08 20:00:00
## 756756     450000 2015-12-28 19:00:00         Q 2010-09-07 20:00:00
##        prior_book_page prior_sale_price editor edit_date collection_year
## 754331            <NA>               NA   <NA>      <NA>            2023
## 756018            <NA>               NA   <NA>      <NA>            2023
## 756488            <NA>               NA   <NA>      <NA>            2023
## 756731       4505/0324                0   <NA>      <NA>            2023
## 756732       4861/0269            83160   <NA>      <NA>            2023
## 756756       8595/0319           450000   <NA>      <NA>            2023
##        planning_region state_use state_use_description
## 754331   South Central      9020          CITY  MDL-94
## 756018   South Central      3530    FRATNL ORG  MDL-94
## 756488   South Central      4300     TEL X STA  MDL-96
## 756731   South Central      4010     IND WHSES  MDL-96
## 756732   South Central      4010     IND WHSES  MDL-96
## 756756   South Central      4000    INDUSTRIAL  MDL-96
##                                      globalid shape_length shape_area sf
## 754331 {147BBAA6-4B6A-4248-B623-1E96D68B76A8}    1055.6852  62010.509  0
## 756018 {7218298A-C821-4D28-8471-369CAE60465D}     632.6066  16165.154  0
## 756488 {EFD5056D-6D04-482F-B5FE-62BB41AB8EF4}     227.9482   2817.456  0
## 756731 {E8034B45-CCFE-4D34-A78F-CB54FBF60624}     463.8206   5836.079  0
## 756732 {53C79594-04D9-419A-B245-36C5BCEC2637}     232.8151   3311.196  0
## 756756 {6B78162F-05FE-4CDE-8474-E0754C3101C6}     194.0787   2292.230  0
##          at_log
## 754331 13.88129
## 756018 13.50276
## 756488 13.66682
## 756731 13.45463
## 756732 12.98691
## 756756 13.03968
```

```
# Remove properties with living areas exceeding 10,000 sq ft
dsf <- dsf[dsf$living_area < 10000 | is.na(dsf$living_area), ]
nrow(dsf)
```

```
## [1] 15296
```

### Effective Year Built (EYB)

EYB data shows inconsistencies; consider removing if it introduces bias.

```
table(dsf$ayb, useNA = 'always')
```

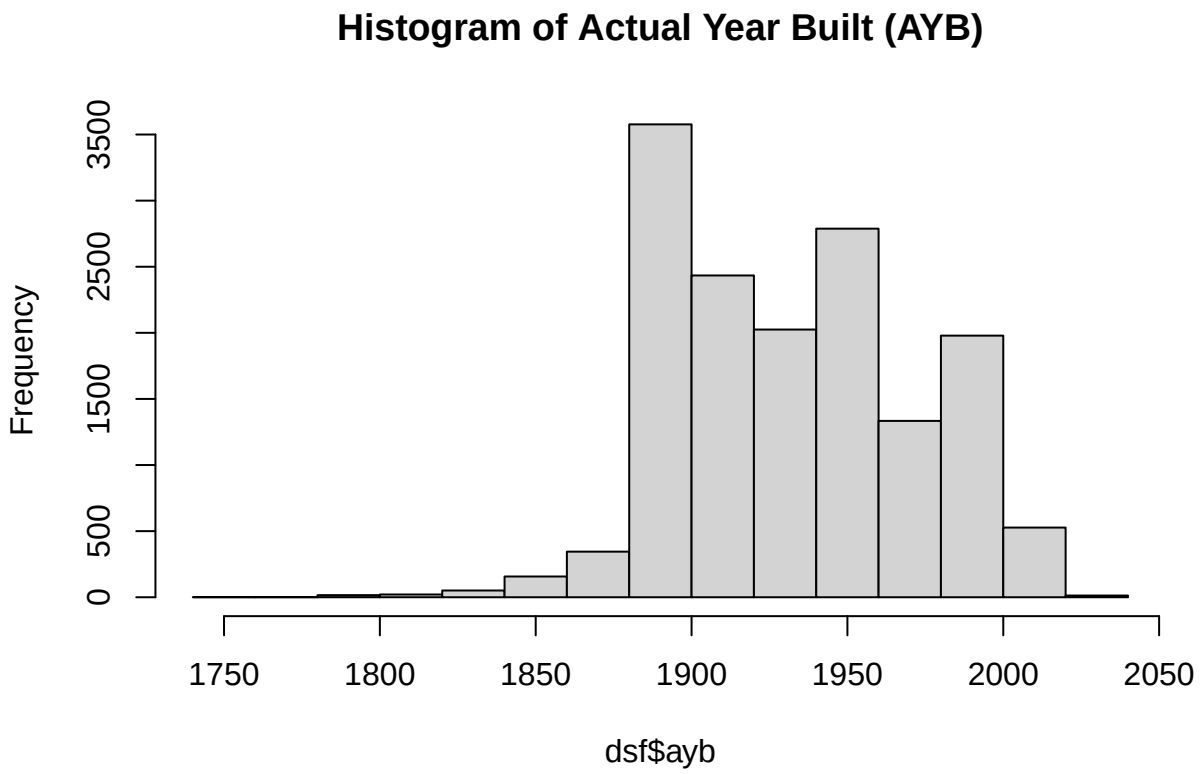
```
##
## 1752 1763 1787 1800 1803 1804 1805 1806 1810 1811 1812 1813 1817 1820 1822 1825
##    1    1    1   15    1    3    1    1    3    1    1    1    1    8    1    4
## 1829 1830 1831 1832 1833 1834 1835 1836 1837 1838 1840 1841 1842 1843 1844 1845
##    1    7    1    7    1    2    2    2    1    5   17    3    1    9    1   10
## 1846 1847 1848 1849 1850 1851 1852 1854 1856 1857 1858 1859 1860 1861 1862 1863
##    8    3    3    5   49    1    3    1    3    7    1    3   46    1    3    1
## 1864 1865 1866 1867 1868 1869 1870 1871 1872 1873 1874 1875 1876 1877 1878 1879
##    3   30    2    1    1    2   87    7    8    2    5   46    9    6    3    7
```

```
## 1880 1882 1883 1884 1885 1886 1887 1888 1889 1890 1891 1892 1893 1894 1895 1896
##  121    6    4    5    8    4    2    7    1  234    6   16    4    3   16    7
## 1897 1898 1899 1900 1901 1902 1903 1904 1905 1906 1907 1908 1909 1910 1911 1912
##    6    5    6 3237    8    7   11    9   58   27   18   18   12  610   20   34
## 1913 1914 1915 1916 1917 1918 1919 1920 1921 1922 1923 1924 1925 1926 1927 1928
##   23   74  160   98   53   82   15 1097   17   29   38   72  238   72  117  120
## 1929 1930 1931 1932 1933 1934 1935 1936 1937 1938 1939 1940 1941 1942 1943 1944
##   67  397   23   32   20   24  152   36   49   70   23  429   47   84   65   38
## 1945 1946 1947 1948 1949 1950 1951 1952 1953 1954 1955 1956 1957 1958 1959 1960
##  153  101  196  166   64  419   90  231   91   93  261  120  138  119   77  235
## 1961 1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974 1975 1976
##  107  117   89  218  102   47   67  136   38   55   20   30   43   71   40   35
## 1977 1978 1979 1980 1981 1982 1983 1984 1985 1986 1987 1988 1989 1990 1991 1992
##   50   13   15   41   50   54   95   98  101  187  476  367  230  136   27   55
## 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008
##   19    6   11   12    6   11   17   21   22   48   52   27   74   68   28   12
## 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 <NA>
##   19   45   22    9   34   12    7    8    7   17    7    9    9    3    1   27
```

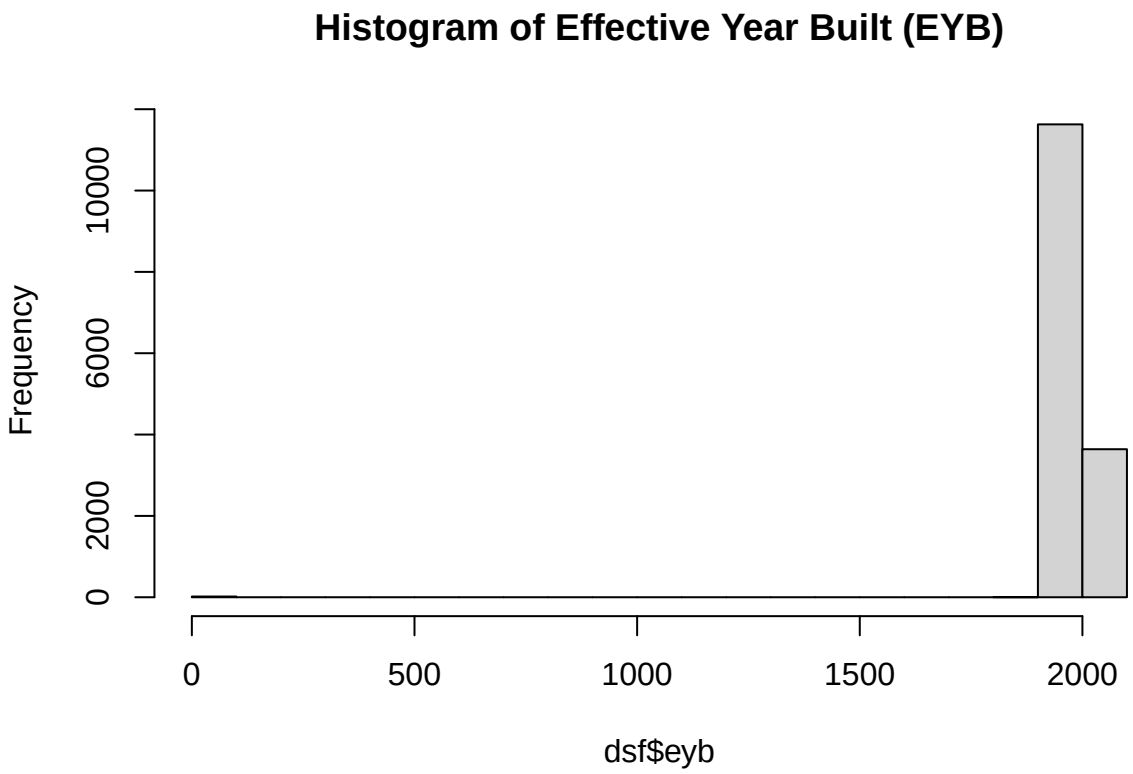
```
table(dsf$eyb, useNA = 'always')
```

```
##
##    0 1900 1943 1958 1961 1969 1971 1972 1976 1979 1981 1984 1985 1986 1987 1988
##   18    1   13   44    4    1   13    1  589    3  670    1    1  467    2    1
## 1989 1991 1992 1993 1994 1995 1996 1997 1999 2000 2001 2002 2003 2004 2005 2006
## 5647  440    3  109  107   15   27   28 3441    1  483    1  218   35  661  119
## 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022
##  176  131  731    2  620   23   81  179   58   40   10   20   26   12   12    1
## 2023 <NA>
##    1    9
```

```
hist(dsf$ayb, main = "Histogram of Actual Year Built (AYB)")
```




```r
hist(dsf$eyb, main = "Histogram of Effective Year Built (EYB)")
```




```
# Check ranges and number of NA values
range(dsf$eyb, na.rm = T)  # eyb is not consistent at all
```

```
## [1]    0 2023
```

```
range(dsf$ayb, na.rm = T)  # ayb has fewer data problems
```

```
## [1] 1752 2023
```

```
sum(is.na(dsf$eyb))
```

```
## [1] 9
```

```
sum(is.na(dsf$ayb))
```

```
## [1] 27
```

We decide to use `AYB` since it is more reliable.

### Zone Descriptions

We categorize zones to understand their effects on housing prices. Zone descriptions and codes are grouped based on zoning regulations using resources like (1) Municode Library and (2) New Haven Zoning and Regulations. In some cases, we also used AI and online searches to categorize zones when direct matches were not available. While this grouping may not be entirely accurate, zoning likely affects housing prices, so it is included in the analysis

```
# Tabulate zone types and descriptions
sort(table(dsf$zone_description, useNA = 'always'))
```

```
##  <NA>
## 15296
```

```
sort(table(dsf$zone, useNA = 'always'))
```

```
##
##   BA/RO     BD3  IL/RM2  PDD 45  PDD 53  PDD 85 PDU 107 PDU 109 PDU 117  PDU 12
##       1       1       1       1       1       1       1       1       1       1
##  PDU 14  PDU 20  PDU 29  PDU 59   PDU 8   PDU 9 RM1/RS2  RM2/RO  BA/RS2 PDU 102
##       1       1       1       1       1       1       1       1       2       2
##  PDU 46  PDU 50  PDU 72 PDU 105 PDU 110  PDU 16  PDU 75 RS1/RS2   PDU 1  PDD 39
##       2       2       2       3       3       3       3       3       5       6
## PDU 108     CEM  IH/RM2    PARK  PDD 49    <NA> PDD 119  PDD 63  PDD 74 RM1/RM2
##       7       8       8       8       8       8      15      16      21      25
## RM2/RS2  PDD 52  PDD 64  BA/RM1  PDD 41 PDU 106  PDD 51  PDD 26  PDD 33      BC
##      25      26      26      27      28      30      31      39      45      52
##     RH2  BA/RM2      BB  PDD 27     BA1  PDD 37      RO      IL  PDD 48  PDD 38
##      53      54      58      62      65      68      71      89      93     103
##      IH      BD     RH1     BD1     RS1      BA     RM1     RM2     RS2
##     131     138     144     149     180     686    3356    4088    5201
```

```
sum(is.na(dsf$zone_description))
```

```
## [1] 15296
```

```
sum(is.na(dsf$zone))
```

```
## [1] 8
```

```
zones <- names(table(dsf$zone))

# 1. Residential Zones
residential_zones <- zones[grep("^RS|RM", zones)]

# 2. Commercial Zones
commercial_zones <- zones[grep("^BA|^BB|^BD", zones)]

# 3. Mixed-Use Zones
mixed_use_zones <- zones[grep("/", zones)]
```

```r
# 4. Industrial Zones
industrial_zones <- zones[grep("^IL|^IH", zones)]

# 5. Planned Development Districts (PDD)
pdd_zones <- zones[grep("^PDD", zones)]

# 6. Public Use Zones (PARK, CEM)
public_use_zones <- zones[grep("PARK|CEM", zones)]

# 7. Specialized Residential/Office Zones
special_res_office_zones <- zones[grep("RO", zones)]

# 8. Parking and Utility Zones (PDU)
parking_utility_zones <- zones[grep("^PDU", zones)]

# 9. Historic and Overlay Zones
historic_overlay_zones <- zones[grep("RH", zones)]

# 10. Undefined/Other Zones
undefined_other_zones <- zones[!zones %in% c(residential_zones, commercial_zones,
                                             mixed_use_zones, industrial_zones,
                                             pdd_zones, public_use_zones,
                                             special_res_office_zones,
                                             parking_utility_zones,
                                             historic_overlay_zones)]

# Reclassify zone variable based on categories
dsf$zone <- ifelse(dsf$zone %in% residential_zones, "Residential", dsf$zone)
dsf$zone <- ifelse(dsf$zone %in% commercial_zones, "Commercial", dsf$zone)
dsf$zone <- ifelse(dsf$zone %in% mixed_use_zones, "Mixed", dsf$zone)
dsf$zone <- ifelse(dsf$zone %in% industrial_zones, "Industrial", dsf$zone)
dsf$zone <- ifelse(dsf$zone %in% pdd_zones, "Planned Development", dsf$zone)
dsf$zone <- ifelse(dsf$zone %in% public_use_zones, "Public Use", dsf$zone)
dsf$zone <- ifelse(dsf$zone %in% special_res_office_zones, "Speical Res", dsf$zone)
dsf$zone <- ifelse(dsf$zone %in% parking_utility_zones, "Parking", dsf$zone)
dsf$zone <- ifelse(dsf$zone %in% historic_overlay_zones, "Historic", dsf$zone)
dsf$zone <- ifelse(dsf$zone %in% undefined_other_zones, "Others", dsf$zone)

# Tabulate the reclassified zones
table(dsf$zone)
```

```
##
##          Commercial            Historic          Industrial              Others
##                1100                 197                 220                  52
##             Parking Planned Development          Public Use         Residential
##                  72                 590                  16               12970
##         Speical Res
##                  71
```

### Condition Description

We inspect the condition descriptions

```
# Inspect condition descriptions
table(dsf$condition_description, useNA = 'always')
```

```
##
##   Average        BA Excellent         F      Good      Poor Very Good Very Poor
##      8405         4       177       608      5341        58       657        18
##      <NA>
##        28
```

```
length(table(dsf$condition_description))
```

```
## [1] 8
```

```
# Reclassify conditions where necessary
dsf$condition_description <- ifelse(dsf$condition_description == "F", "Fair",
                                    dsf$condition_description)


# Does "BA" mean bad? NO
# Checked properties to confirm
# Checked 46 CLINTON AV: warehouse
# Checked 102 FOOD TERMINAL PLZ: warehouse
head(dsf[dsf$condition_description == "BA" &
           !is.na(dsf$condition_description), ])
```

```
##        town_name                 link                         owner
## 763627 NEW HAVEN 52070-161 0760 01300             Y&S PROPERTIES LLC
## 770032 NEW HAVEN 52070-236 1304 00701              PETRILLO RONALD A
## 774198 NEW HAVEN 52070-299 0144 02900              SANTOS ALBERTO SR
## 776700 NEW HAVEN 52070-338 0209 00700 CHARISMATIC RENEWAL MINISTRIES
##                        co_owner              location           mailing_address
## 763627                     <NA>         46 CLINTON AV 91 SHELTON AVE, SUITE 112
## 770032                     <NA> 102 FOOD TERMINAL PLZ          518 RACEBROOK RD
## 774198 VELOZ JOHNNY EMMANUEL SR     251 DAVENPORT AV          251 DAVENPORT AV
## 776700                  INC USA          16 NORTON ST               PO BOX 4009
##        mailing_city mailing_state assessed_total assessed_land
## 763627    NEW HAVEN            CT         179900         50330
## 770032       ORANGE            CT         333760        213500
## 774198    NEW HAVEN            CT         271600         28980
## 776700    NEW HAVEN            CT         230090         60130
##        assessed_building pre_year_assessed_total appraised_land
## 763627            121450                  179900          71900
## 770032            119910                  333760         305000
## 774198            242620                  271600          41400
## 776700            165340                  230090          85900
##        appraised_building appraised_outbuilding appraised_extra_feature
## 763627             173500                     0                      NA
## 770032             171300                     0                      NA
## 774198             346600                     0                      NA
## 776700             236200                  6600                      NA
##        valuation_year        zone zone_description model condition
## 763627           2021 Residential             <NA>    94        BA
```

```
## 770032           2021  Industrial             <NA>    96        BA
## 774198           2021 Residential             <NA>    94        BA
## 776700           2021  Commercial             <NA>    94        BA
##        condition_description  ayb  eyb living_area effective_area total_rooms
## 763627                    BA 1920 1979        6626           6630          NA
## 770032                    BA 1979 1987        4713           4722          NA
## 774198                    BA 1900 1979        6125           6475          NA
## 776700                    BA 1910 1979        2587           2847          NA
##        number_of_bedroom number_of_baths number_of_half_baths occupancy
## 763627                NA              NA                   NA         2
## 770032                NA              NA                   NA         1
## 774198                NA              NA                   NA         3
## 776700                NA              NA                   NA         2
##        sale_price           sale_date qualified     prior_sale_date
## 763627     185000 2021-01-11 19:00:00         U 1999-04-04 20:00:00
## 770032     158500 2003-06-18 20:00:00         U 1989-08-31 20:00:00
## 774198     370000 2021-04-14 20:00:00         Q 2018-11-25 19:00:00
## 776700      85000 2011-08-17 20:00:00         U 2011-03-08 19:00:00
##        prior_book_page prior_sale_price editor edit_date collection_year
## 763627       5476/0326            87698   <NA>      <NA>            2023
## 770032       4138/0327                0   <NA>      <NA>            2023
## 774198       9788/0277           205000   <NA>      <NA>            2023
## 776700       8667/0067           101500   <NA>      <NA>            2023
##        planning_region state_use state_use_description
## 763627   South Central      3160     COMM WHSE  MDL-94
## 770032   South Central      4010     IND WHSES  MDL-96
## 774198   South Central      3030     MIXED USE  MDL-94
## 776700   South Central      9060        CHURCH  MDL-94
##                                    globalid shape_length shape_area sf
## 763627 {DED9CAF8-A762-4E63-AC37-E1D3AE52F007}    193.40838  2008.2655  0
## 770032 {C3066D7C-234A-4F87-801D-393830B9E105}    343.59578  5730.8708  0
## 774198 {54FA29FA-07E6-43B2-8F5D-1F768C52081C}     81.22277   387.0876  0
## 776700 {BFEF8208-CC7C-4C09-8DDF-3FB0D7196507}    146.35972   973.7525  0
##          at_log
## 763627 12.10016
## 770032 12.71818
## 774198 12.51209
## 776700 12.34623
```

```
# Does F mean Fair? YES
# Does "F" mean Fair? Checked properties to confirm
# Checked 150 HUNTINGTON RD: fair condition (our judgement)
# Checked 204 BURR ST: fair condition (our judgement)
# Checked 351 CONCORD ST: fair condition (our judgement)
head(dsf[dsf$condition_description == "Fair"
         & !is.na(dsf$condition_description), ])
```

```
##        town_name                link                                owner
## 754657 NEW HAVEN 52070-023 0927 00800                    VAZQUEZ MILAGROS
## 754899 NEW HAVEN 52070-027 0905 00600                GAUDINO RITA & JASON T
## 754975 NEW HAVEN 52070-028 0895 00100 CITY OF NEW HAVEN HOUSING AUTHORITY
## 755056 NEW HAVEN 52070-028 0898 03100          FAREZ MARCO FABIAN SANCHEZ
## 755075 NEW HAVEN 52070-029 0889 00100                        CASEY REBECCA
## 755130 NEW HAVEN 52070-029 0892 00700                        FIENGO ROBERT
```

```
##                co_owner           location    mailing_address mailing_city
## 754657       LEON EFRAIN 150 HUNTINGTON RD  150 HUNTINGTON RD     NEW HAVEN
## 754899              <NA>        204 BURR ST        204 BURR ST    NEW HAVEN
## 754975 CITY OF NEW HAVEN     351 CONCORD ST       P O BOX 1912    NEW HAVEN
## 755056              <NA>         90 LEY ST          90 LEY ST    NEW HAVEN
## 755075              <NA>     257 CONCORD ST     257 CONCORD ST    NEW HAVEN
## 755130              <NA>          12 IRA ST          12 IRA ST    NEW HAVEN
##        mailing_state assessed_total assessed_land assessed_building
## 754657            CT         177730         64820             111090
## 754899            CT         151200         56140              93240
## 754975            CT         129850         62370              67480
## 755056            CT         156730         73290              77770
## 755075            CT         116270         62230              46970
## 755130            CT         117180         62930              53550
##        pre_year_assessed_total appraised_land appraised_building
## 754657                  177730          92600              158700
## 754899                  151200          80200              133200
## 754975                  129850          89100               96400
## 755056                  156730         104700              111100
## 755075                  116270          88900               67100
## 755130                  117180          89900               76500
##        appraised_outbuilding appraised_extra_feature valuation_year        zone
## 754657                     0                      NA           2021 Residential
## 754899                     0                      NA           2021 Residential
## 754975                     0                      NA           2021 Residential
## 755056                  8100                      NA           2021 Residential
## 755075                 10100                      NA           2021 Residential
## 755130                  1000                      NA           2021 Residential
##        zone_description model condition condition_description  ayb  eyb
## 754657             <NA>     1         F                  Fair 1964 1976
## 754899             <NA>     1         F                  Fair 1952 1976
## 754975             <NA>     1         F                  Fair 1900 1976
## 755056             <NA>     1         F                  Fair 1947 1976
## 755075             <NA>     1         F                  Fair 1900 1976
## 755130             <NA>     1         F                  Fair 1955 1976
##        living_area effective_area total_rooms number_of_bedroom number_of_baths
## 754657        1536           2059           6                 3               2
## 754899        1358           1842           6                 3               1
## 754975        1368           1520           6                 3               1
## 755056        1476           1663           5                 4               1
## 755075         768            944           6                 2               1
## 755130         800            984           4                 2               2
##        number_of_half_baths occupancy sale_price           sale_date qualified
## 754657                    1         1     230000 2021-01-24 19:00:00         Q
## 754899                    0         1          0 2018-08-06 20:00:00         U
## 754975                    0         1     100000 1993-01-13 19:00:00         Q
## 755056                    1         1     199000 2020-10-14 20:00:00         Q
## 755075                    0         1     120000 2015-09-14 20:00:00         Q
## 755130                    0         1          0 2012-07-17 20:00:00         U
##            prior_sale_date prior_book_page prior_sale_price editor edit_date
## 754657 2013-11-25 19:00:00       9082/0272           235000   <NA>      <NA>
## 754899 2015-01-29 19:00:00       9242/0178                0   <NA>      <NA>
## 754975                <NA>            <NA>               NA   <NA>      <NA>
## 755056 2017-09-17 20:00:00       9623/0156           152000   <NA>      <NA>
```

```
## 755075 2012-06-11 20:00:00       8842/0273                0   <NA>      <NA>
## 755130 2005-06-27 20:00:00       7236/0304                0   <NA>      <NA>
##        collection_year planning_region state_use state_use_description
## 754657            2023   South Central      1010         Single Family
## 754899            2023   South Central      1010         Single Family
## 754975            2023   South Central      <NA>     HSNG AUTH  MDL-01
## 755056            2023   South Central      1010         Single Family
## 755075            2023   South Central      1010         Single Family
## 755130            2023   South Central      1010         Single Family
##                                      globalid shape_length shape_area sf
## 754657 {F7C01077-237F-4A5E-8A89-7E30097A150B}     197.6596   2052.702  1
## 754899 {86F4AB61-946E-4811-8365-AA53511A8668}     169.0062   1396.023  1
## 754975 {8ACD32AC-18CE-4954-AE74-3E66062AA513}     140.5247   1006.158  0
## 755056 {5C60009A-CF7D-494A-928D-EDC6668DB9AE}     260.3038   3553.854  1
## 755075 {9C197826-C4E1-4845-B7E2-AEE33D280EA4}     143.4300   1040.854  1
## 755130 {4D3C3E01-1FF0-402B-B726-DE9964687DFB}     136.9891   1004.754  1
##          at_log
## 754657 12.08802
## 754899 11.92636
## 754975 11.77414
## 755056 11.96228
## 755075 11.66367
## 755130 11.67147
```

```
# Remove entries with "BA" as it doesn't match our criteria
dsf <- dsf[dsf$condition_description != "BA" |
             is.na(dsf$condition_description), ]

# Convert condition_description to a factor variable
dsf$condition_description <- factor(dsf$condition_description,
                                    levels = c("Very Poor", "Poor", "Fair",
                                               "Average", "Good", "Very Good",
                                               "Excellent"))

# Verify that the factor levels are set correctly
levels(dsf$condition_description)
```

```
## [1] "Very Poor" "Poor"      "Fair"      "Average"   "Good"      "Very Good"
## [7] "Excellent"
```

```
# Assess missing values in condition_description
nrow(dsf[is.na(dsf$condition_description), ])
```

```
## [1] 28
```

```
head(dsf[is.na(dsf$condition_description), ])
```

```
##        town_name                 link                                  owner
## 757244 NEW HAVEN 52070-076 0988 01100          CITY OF NEW HAVEN HOUSING AUTH
## 758452 NEW HAVEN 52070-098 1012 00300                     PUTNAM COVE LLC
## 760402 NEW HAVEN 52070-181 0589 00400                   BRUCKUF ANGELIKA D
## 761601 NEW HAVEN 52070-021 0919 00801                    CITY OF NEW HAVEN
```

```
## 761651 NEW HAVEN 52070-066 0951 00400                NEW HAVEN PORT AUTHORITY
## 761652 NEW HAVEN 52070-052 0950 00500 GREATER NEW HAVEN WATER POLLUTION CONTRO
##                co_owner          location        mailing_address mailing_city
## 757244 CITY OF NEW HAVEN   70 FAIRMONT AV           360 ORANGE ST    NEW HAVEN
## 758452              <NA> 16 EAST GRAND AV          94 CLEMENTS RD       NEWTON
## 760402              <NA> 21 MILL RIVER ST        27 MILL RIVER ST    NEW HAVEN
## 761601              <NA>           BURR ST          165 CHURCH ST    NEW HAVEN
## 761651              <NA>    CONNECTICUT AV          200 ORANGE ST    NEW HAVEN
## 761652              <NA>    CONNECTICUT AV 345 EAST SHORE PARKWAY    NEW HAVEN
##        mailing_state assessed_total assessed_land assessed_building
## 757244            CT        1089620       1089620                0
## 758452            MA         234010        234010                0
## 760402            CT          81690         69860                0
## 761601            CT          54880         54880                0
## 761651            CT        1907430       1907430                0
## 761652            CT         171010        171010                0
##        pre_year_assessed_total appraised_land appraised_building
## 757244                 1089620        1556600                  0
## 758452                  276850         334300                  0
## 760402                   81690          99800                  0
## 761601                   54880          78400                  0
## 761651                 1907430        2724900                  0
## 761652                  171010         244300                  0
##        appraised_outbuilding appraised_extra_feature valuation_year        zone
## 757244                     0                      NA           2021 Residential
## 758452                     0                      NA           2021     Parking
## 760402                 16900                      NA           2021 Residential
## 761601                     0                      NA           2021 Residential
## 761651                     0                      NA           2021  Industrial
## 761652                     0                      NA           2021  Public Use
##        zone_description model condition condition_description ayb eyb
## 757244             <NA>     0      <NA>                 <NA>  NA  NA
## 758452             <NA>     0      <NA>                 <NA>  NA  NA
## 760402             <NA>     0      <NA>                 <NA>  NA  NA
## 761601             <NA>     0      <NA>                 <NA>  NA   0
## 761651             <NA>     0      <NA>                 <NA>  NA   0
## 761652             <NA>     0      <NA>                 <NA>  NA   0
##        living_area effective_area total_rooms number_of_bedroom number_of_baths
## 757244          NA             NA          NA                NA              NA
## 758452          NA             NA          NA                NA              NA
## 760402          NA             NA          NA                NA              NA
## 761601          NA             NA          NA                NA              NA
## 761651          NA             NA          NA                NA              NA
## 761652          NA             NA          NA                NA              NA
##        number_of_half_baths occupancy sale_price           sale_date qualified
## 757244                   NA        NA          0 1981-05-11 20:00:00         U
## 758452                   NA        NA          0 2016-02-28 19:00:00         U
## 760402                   NA        NA          0 2018-09-26 20:00:00         U
## 761601                   NA        NA          0 2011-08-17 20:00:00         U
## 761651                   NA        NA          0 2008-03-06 19:00:00         U
## 761652                   NA        NA          0 2008-03-06 19:00:00         U
##            prior_sale_date prior_book_page prior_sale_price editor edit_date
## 757244                <NA>            <NA>               NA   <NA>      <NA>
## 758452 2015-06-21 20:00:00       9296/0079                0   <NA>      <NA>
```

```
## 760402 2017-10-04 20:00:00       9629/0176               0   <NA>      <NA>
## 761601 2006-07-12 20:00:00       7647/0163               0   <NA>      <NA>
## 761651               <NA>            <NA>              NA   <NA>      <NA>
## 761652               <NA>            <NA>              NA   <NA>      <NA>
##        collection_year planning_region state_use state_use_description
## 757244            2023   South Central      <NA>     HSNG AUTH  MDL-00
## 758452            2023   South Central      3900            DEVEL LAND
## 760402            2023   South Central      3900            DEVEL LAND
## 761601            2023   South Central      <NA>          CITY  MDL-00
## 761651            2023   South Central      <NA>     MUNICIPAL  MDL-00
## 761652            2023   South Central      <NA>     MUNICIPAL  MDL-00
##                                     globalid shape_length shape_area sf
## 757244 {89E45E1F-AD19-4BA0-B45D-32B38C76B837}     516.3895  16239.308  0
## 758452 {8663B825-2617-44AF-A605-4AD1A6FD6A11}     304.9546   5736.941  0
## 760402 {FBD72C54-A9CB-4E07-8E72-0622A5F82394}     157.9034   1455.939  0
## 761601 {0814A65A-5F69-4040-8B0A-1D15F9FEBF20}     194.1900   1202.413  0
## 761651 {C4AC87D2-4A17-407B-AC53-E3ADC262CFBC}    2347.9373  78988.046  0
## 761652 {6F84018B-E5B4-4D09-A802-936B14A80E8C}     318.4276   6066.026  0
##          at_log
## 757244 13.90134
## 758452 12.36312
## 760402 11.31069
## 761601 10.91290
## 761651 14.46127
## 761652 12.04948
```

## STEP (4) Final Check:

Assessing and Understanding Data Quality

```
# Condition Description
unique(dsf$condition_description)
```

```
## [1] Good      Average   Excellent Very Good Fair      Poor      Very Poor
## [8] <NA>
## Levels: Very Poor Poor Fair Average Good Very Good Excellent
```

```
sum(is.na(dsf$condition_description))
```

```
## [1] 28
```

```
table(dsf$condition_description, useNA = "always")
```

```
##
## Very Poor      Poor      Fair   Average      Good Very Good Excellent      <NA>
##        18        58       608      8405      5341       657       177        28
```

```
head(dsf[is.na(dsf$condition_description), ])
```

```
##        town_name                link                                    owner
```

```
## 757244 NEW HAVEN 52070-076 0988 01100          CITY OF NEW HAVEN HOUSING AUTH
## 758452 NEW HAVEN 52070-098 1012 00300                      PUTNAM COVE LLC
## 760402 NEW HAVEN 52070-181 0589 00400                   BRUCKUF ANGELIKA D
## 761601 NEW HAVEN 52070-021 0919 00801                    CITY OF NEW HAVEN
## 761651 NEW HAVEN 52070-066 0951 00400                NEW HAVEN PORT AUTHORITY
## 761652 NEW HAVEN 52070-052 0950 00500 GREATER NEW HAVEN WATER POLLUTION CONTRO
##                co_owner          location        mailing_address mailing_city
## 757244 CITY OF NEW HAVEN   70 FAIRMONT AV          360 ORANGE ST    NEW HAVEN
## 758452              <NA> 16 EAST GRAND AV          94 CLEMENTS RD       NEWTON
## 760402              <NA> 21 MILL RIVER ST        27 MILL RIVER ST    NEW HAVEN
## 761601              <NA>           BURR ST          165 CHURCH ST    NEW HAVEN
## 761651              <NA>    CONNECTICUT AV          200 ORANGE ST    NEW HAVEN
## 761652              <NA>    CONNECTICUT AV 345 EAST SHORE PARKWAY    NEW HAVEN
##        mailing_state assessed_total assessed_land assessed_building
## 757244            CT        1089620       1089620                 0
## 758452            MA         234010        234010                 0
## 760402            CT          81690         69860                 0
## 761601            CT          54880         54880                 0
## 761651            CT        1907430       1907430                 0
## 761652            CT         171010        171010                 0
##        pre_year_assessed_total appraised_land appraised_building
## 757244                 1089620        1556600                  0
## 758452                  276850         334300                  0
## 760402                   81690          99800                  0
## 761601                   54880          78400                  0
## 761651                 1907430        2724900                  0
## 761652                  171010         244300                  0
##        appraised_outbuilding appraised_extra_feature valuation_year        zone
## 757244                     0                      NA           2021 Residential
## 758452                     0                      NA           2021     Parking
## 760402                 16900                      NA           2021 Residential
## 761601                     0                      NA           2021 Residential
## 761651                     0                      NA           2021  Industrial
## 761652                     0                      NA           2021  Public Use
##        zone_description model condition condition_description ayb eyb
## 757244             <NA>     0      <NA>                  <NA>  NA  NA
## 758452             <NA>     0      <NA>                  <NA>  NA  NA
## 760402             <NA>     0      <NA>                  <NA>  NA  NA
## 761601             <NA>     0      <NA>                  <NA>  NA   0
## 761651             <NA>     0      <NA>                  <NA>  NA   0
## 761652             <NA>     0      <NA>                  <NA>  NA   0
##        living_area effective_area total_rooms number_of_bedroom number_of_baths
## 757244          NA             NA          NA                NA              NA
## 758452          NA             NA          NA                NA              NA
## 760402          NA             NA          NA                NA              NA
## 761601          NA             NA          NA                NA              NA
## 761651          NA             NA          NA                NA              NA
## 761652          NA             NA          NA                NA              NA
##        number_of_half_baths occupancy sale_price           sale_date qualified
## 757244                   NA        NA          0 1981-05-11 20:00:00         U
## 758452                   NA        NA          0 2016-02-28 19:00:00         U
## 760402                   NA        NA          0 2018-09-26 20:00:00         U
## 761601                   NA        NA          0 2011-08-17 20:00:00         U
## 761651                   NA        NA          0 2008-03-06 19:00:00         U
```

```
## 761652                   NA        NA          0 2008-03-06 19:00:00         U
##            prior_sale_date prior_book_page prior_sale_price editor edit_date
## 757244                <NA>            <NA>               NA   <NA>      <NA>
## 758452 2015-06-21 20:00:00       9296/0079                0   <NA>      <NA>
## 760402 2017-10-04 20:00:00       9629/0176                0   <NA>      <NA>
## 761601 2006-07-12 20:00:00       7647/0163                0   <NA>      <NA>
## 761651                <NA>            <NA>               NA   <NA>      <NA>
## 761652                <NA>            <NA>               NA   <NA>      <NA>
##        collection_year planning_region state_use state_use_description
## 757244            2023   South Central      <NA>     HSNG AUTH  MDL-00
## 758452            2023   South Central      3900            DEVEL LAND
## 760402            2023   South Central      3900            DEVEL LAND
## 761601            2023   South Central      <NA>          CITY  MDL-00
## 761651            2023   South Central      <NA>     MUNICIPAL  MDL-00
## 761652            2023   South Central      <NA>     MUNICIPAL  MDL-00
##                                     globalid shape_length shape_area sf
## 757244 {89E45E1F-AD19-4BA0-B45D-32B38C76B837}     516.3895  16239.308  0
## 758452 {8663B825-2617-44AF-A605-4AD1A6FD6A11}     304.9546   5736.941  0
## 760402 {FBD72C54-A9CB-4E07-8E72-0622A5F82394}     157.9034   1455.939  0
## 761601 {0814A65A-5F69-4040-8B0A-1D15F9FEBF20}     194.1900   1202.413  0
## 761651 {C4AC87D2-4A17-407B-AC53-E3ADC262CFBC}    2347.9373  78988.046  0
## 761652 {6F84018B-E5B4-4D09-A802-936B14A80E8C}     318.4276   6066.026  0
##          at_log
## 757244 13.90134
## 758452 12.36312
## 760402 11.31069
## 761601 10.91290
## 761651 14.46127
## 761652 12.04948
```

```
# Number of Rooms: Bedrooms, Bathrooms, Half Baths
sum(is.na(dsf$total_rooms))
```

```
## [1] 1712
```

```
sum(is.na(dsf$number_of_bedroom))
```

```
## [1] 1617
```

```
sum(is.na(dsf$number_of_baths))
```

```
## [1] 1617
```

```
sum(is.na(dsf$number_of_halfbaths))
```

```
## [1] 0
```

```
# Zone
sum(is.na(dsf$zone))
```

```
## [1] 8
```

```r
# Living Area
sum(is.na(dsf$living_area))
```

```
## [1] 27
```

```r
# Assessed Total
sum(is.na(dsf$assessed_total))
```

```
## [1] 0
```

### STEP(5): Select Variables for Analysis and Save the data

Select key variables for analysis to simplify the data processing

```r
# Key Variables of Interests
vars_interests <- c(
  "at_log", "assessed_total", "zone",
  "condition_description", "ayb", "total_rooms",
  "number_of_bedroom", "number_of_baths",
  "living_area"
)


# Key Variables of Distance
dist_vars <- c(
  "dist_convenience_store", "dist_farmers_and_markets",
  "dist_grocery_store", "dist_super_store",
  "dist_supermarket", "area_super_store",
  "area_supermarket", "area_grocery_store"
)

# Select relevant columns for analysis
dr <- dsf[, vars_interests]
saveRDS(dr, "cleaned.rds")
```

## Conclusion

In this file, we have cleaned the dataframe

## Sessioninfo

```
sessionInfo()
```

```
## R version 4.4.1 (2024-06-14)
## Platform: aarch64-apple-darwin20
## Running under: macOS 15.0.1
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK v
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: America/New_York
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] sf_1.0-17
##
## loaded via a namespace (and not attached):
##  [1] digest_0.6.37      fastmap_1.2.0      xfun_0.47          magrittr_2.0.3
##  [5] e1071_1.7-14       KernSmooth_2.23-24 knitr_1.48         htmltools_0.5.8.1
##  [9] rmarkdown_2.28     classInt_0.4-10    cli_3.6.3          grid_4.4.1
## [13] DBI_1.2.3          proxy_0.4-27       class_7.3-22       compiler_4.4.1
## [17] highr_0.11         rstudioapi_0.16.0  tools_4.4.1        evaluate_0.24.0
## [21] Rcpp_1.0.13        yaml_2.3.10        rlang_1.1.4        units_0.8-5
```

# Appendix (Other information)

**Description of Zones:**

I found the definitions and zoning regulations for New Haven, CT, which will help categorize the zones into about 10 categories. Here's an overview:

1. **Residential Zones (RS, RM)**: These zones include various types of residential areas like single-family (RS1, RS2) and multi-family (RM1, RM2). They typically differ based on density and the type of dwellings allowed.
2. **Commercial Zones (BA, BB, BD)**: These include zones designated for businesses and commercial activities (e.g., BA, BD, BB). The differences often relate to the scale and type of commercial development permitted.
3. **Mixed-Use Zones (e.g., BA/RM1, RM2/RO)**: These zones allow a combination of residential and commercial or other uses, promoting development that integrates living, working, and recreational spaces.
4. **Industrial Zones (IL, IH)**: These zones are for light (IL) and heavy (IH) industrial activities. They regulate the types of manufacturing and industrial processes allowed.
5. **Planned Development Districts (PDD)**: These are special districts (e.g., PDD 45, PDD 53) created for specific development projects or areas that require unique regulations not covered by standard zoning categories.
6. **Public Use Zones (PARK, CEM)**: These zones are designated for public amenities like parks and cemeteries, preserving open spaces for recreational and community purposes.
7. **Historic and Overlay Zones**: Some areas may have historic designations or specific overlays that impose additional regulations to maintain certain aesthetic or architectural standards.
8. **Specialized Residential/Office Zones (RO)**: These zones accommodate both residential and office spaces, providing flexibility in usage depending on the location and needs of the area.
9. **Parking and Utility Zones (PDU)**: Zones such as PDU are designated for parking or specific utility-related uses, ensuring infrastructure support for residential and commercial areas.
10. **Unique or Undefined Categories**: Some combinations like BA/RO or RM1/RS2 may combine characteristics of multiple zones. They are typically used to allow flexibility in transitioning areas.

Reference Resources - Municode. - New Haven Zoning and Regulations