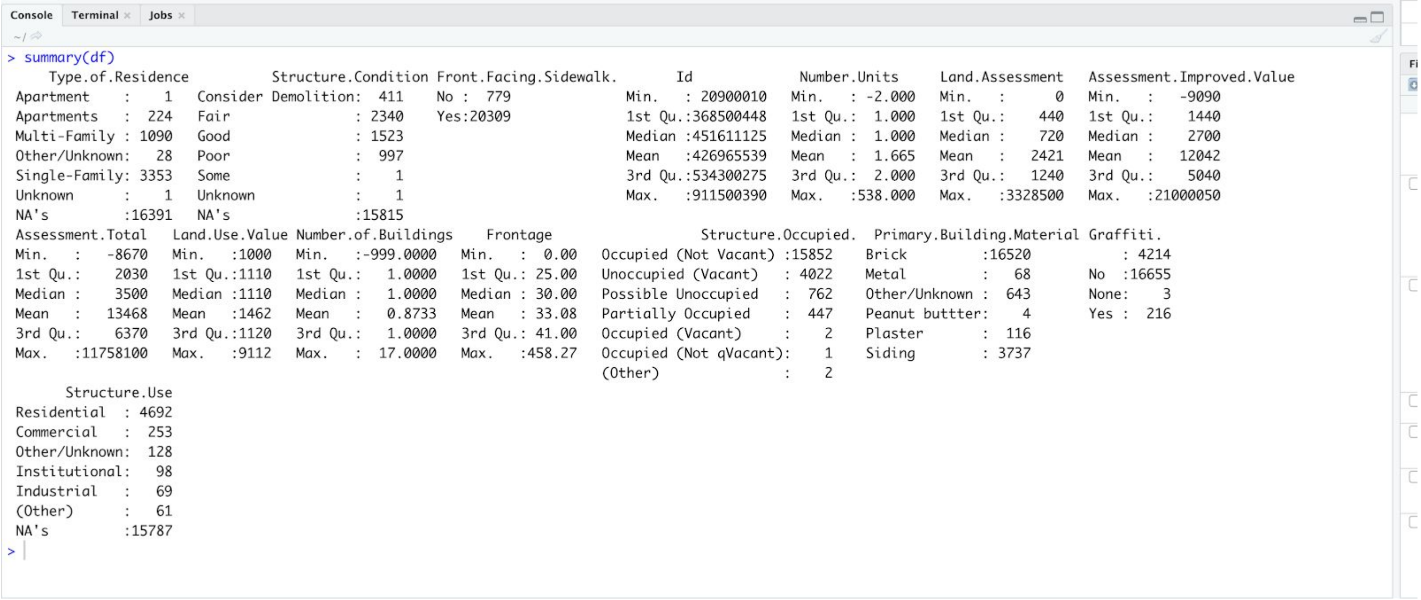
James Valles
Hmk 1: DSC 441

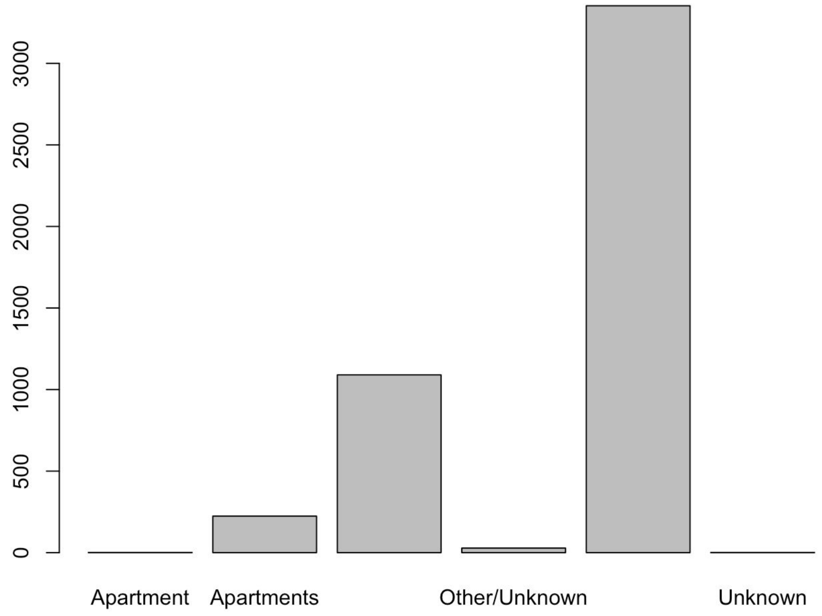


```
> summary(df)
     Type.of.Residence          Structure.Condition Front.Facing.Sidewalk.       Id             Number.Units     Land.Assessment    Assessment.Improved.Value
 Apartment    :    1   Consider Demolition:  411    No :  779              Min.   : 20900010   Min.   : -2.000   Min.   :      0   Min.   :   -9090  
 Apartments   :  224   Fair               : 2340    Yes:20309              1st Qu.:368500448   1st Qu.:  1.000   1st Qu.:    440   1st Qu.:    1440  
 Multi-Family : 1090   Good               : 1523                           Median :451611125   Median :  1.000   Median :    720   Median :    2700  
 Other/Unknown:   28   Poor               :  997                           Mean   :426965539   Mean   :  1.665   Mean   :   2421   Mean   :   12042  
 Single-Family: 3353   Some               :    1                           3rd Qu.:534300275   3rd Qu.:  2.000   3rd Qu.:   1240   3rd Qu.:    5040  
 Unknown      :    1   Unknown            :    1                           Max.   :911500390   Max.   :538.000   Max.   :3328500   Max.   :21000050  
 NA's         :16391   NA's               :15815                                                                                                       
 Assessment.Total   Land.Use.Value Number.of.Buildings    Frontage                   Structure.Occupied.  Primary.Building.Material Graffiti.
 Min.   :   -8670   Min.   :1000   Min.   :-999.0000   Min.   :  0.00   Occupied (Not Vacant) :15852    Brick         :16520          : 4214  
 1st Qu.:    2030   1st Qu.:1110   1st Qu.:   1.0000   1st Qu.: 25.00   Unoccupied (Vacant)   : 4022    Metal         :   68       No   :16655  
 Median :    3500   Median :1110   Median :   1.0000   Median : 30.00   Possible Unoccupied   :  762    Other/Unknown :  643       None:    3  
 Mean   :   13468   Mean   :1462   Mean   :   0.8733   Mean   : 33.08   Partially Occupied    :  447    Peanut buttter:    4       Yes :  216  
 3rd Qu.:    6370   3rd Qu.:1120   3rd Qu.:   1.0000   3rd Qu.: 41.00   Occupied (Vacant)     :    2    Plaster       :  116                   
 Max.   :11758100   Max.   :9112   Max.   :  17.0000   Max.   :458.27   Occupied (Not qVacant):    1    Siding        : 3737                   
                                                                        (Other)               :    2                                             
       Structure.Use 
 Residential  : 4692  
 Commercial   :  253  
 Other/Unknown:  128  
 Institutional:   98  
 Industrial   :   69  
 (Other)      :   61  
 NA's         :15787  
> 
```

First thing I did to answer questions 1-4, I did a summary of the data using R command: summary(df). This provided a ton of insight into missing data values, data type/data formatting errors, as well as, values that were not plausible.

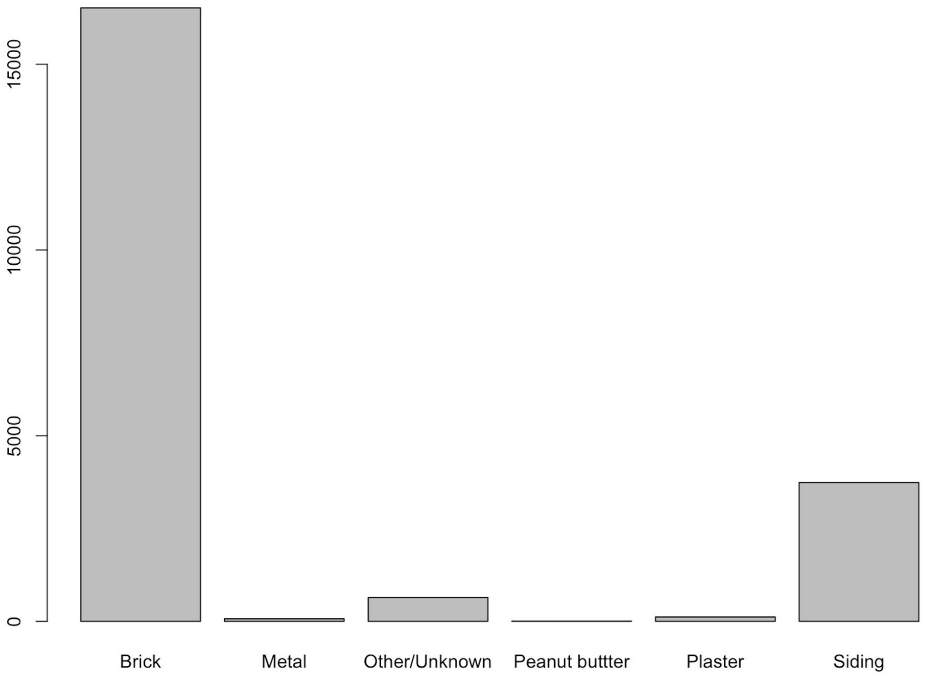
1. **Data Consistency problems:**

Command: **summary(df)** (screenshot above)
For, **Type.of.Residence**, there are inconsistent values: such as 1 entry shows ‘Apartment’ and 224 are for ‘Apartments’. ‘Apartment’ and ‘Apartments’ are essentially the same thing, but they can be grouped into one.

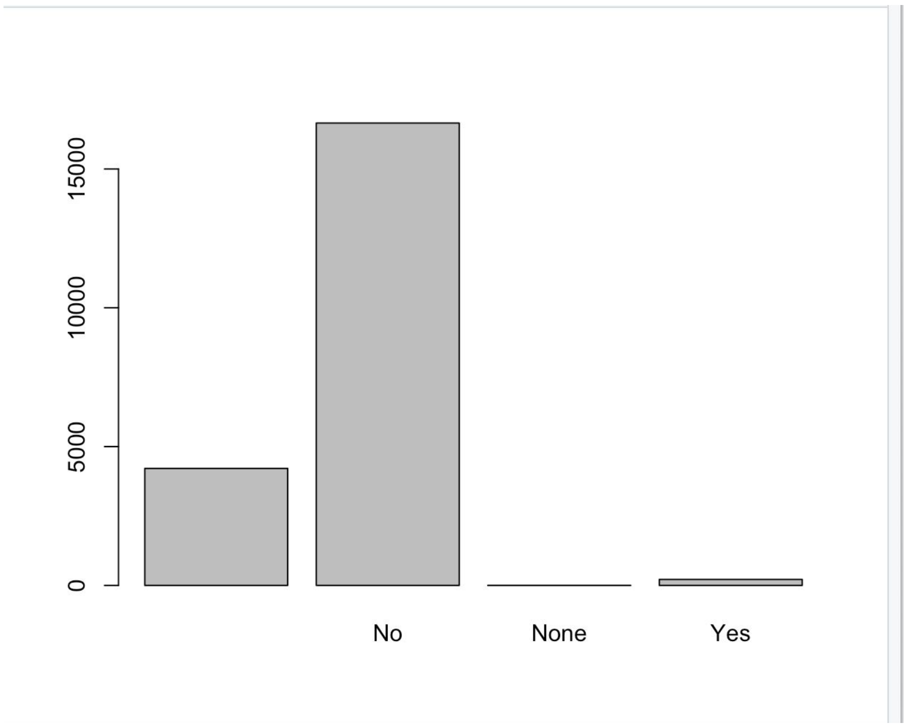


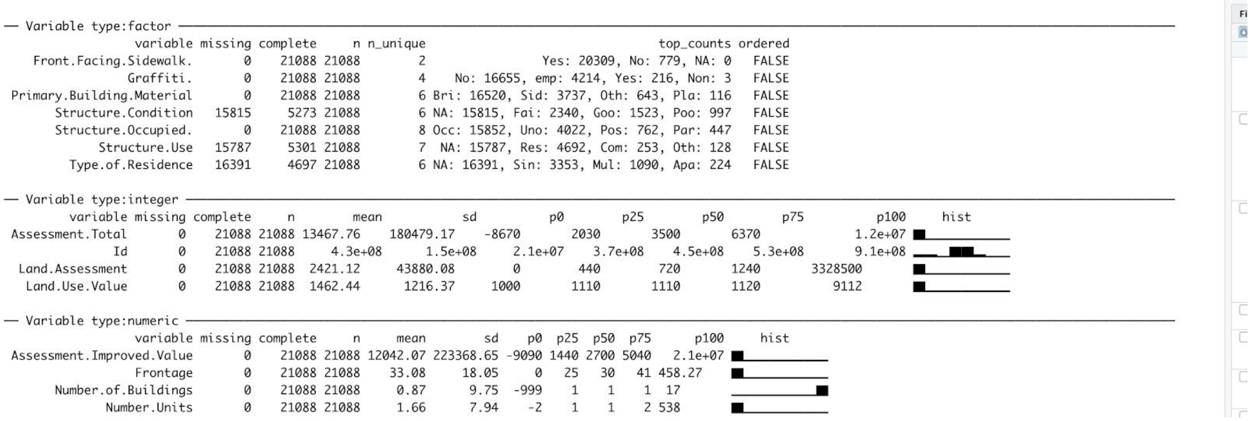
Additionally, I noticed there are two spots for Unknown. Unknown has 1 entry and Other/Unknown has 28. Same thing with **Structure.Use** has Other/Unknown 128 rows, (Other), which has parenthesis wrapped around the word Other, has 61 rows.

One of the most interesting data consistency which isn't plausible is **Primary.Building.Material** has 4 entries for 'Peanut butter,' which is not consistent with what is normally used to build a structure and isn't plausible either.



Also, **Graffiti.** had what appears to be a data consistency and data completeness problem. It had 4,214 rows that had a empty string "" as a value, which is not consistent with the other values. And, had two options for having no Graffiti - 'No' and 'None,' which mean the same thing.



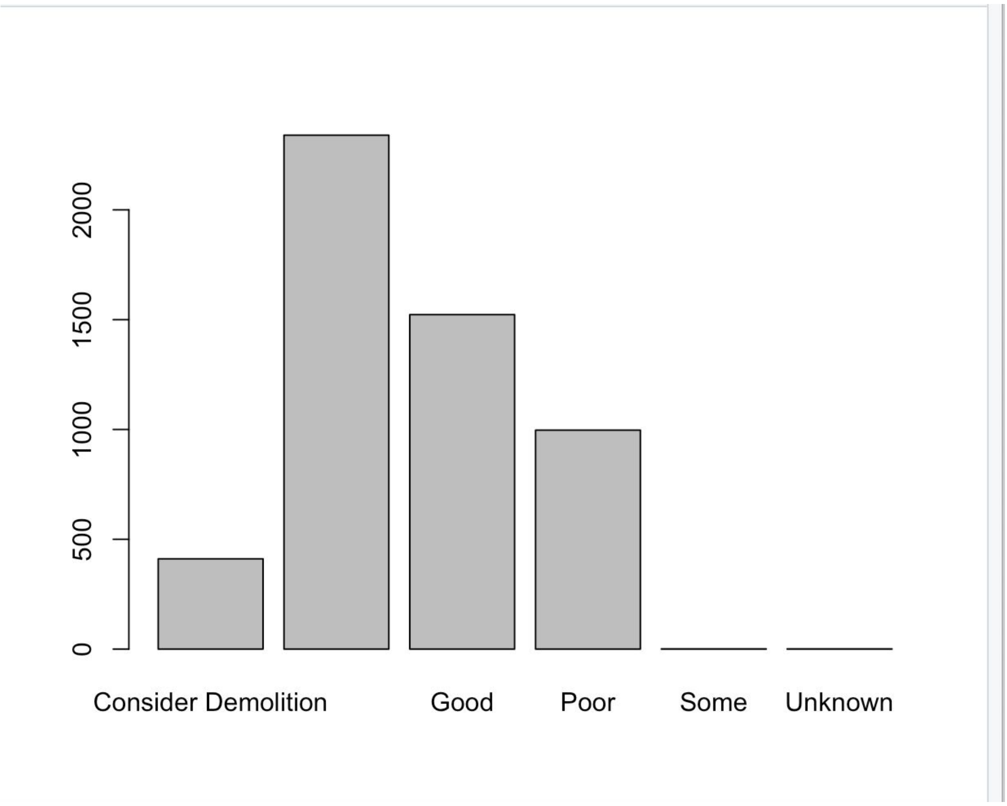


```
— Variable type:factor ————————————————————————————————————————————————————————————————————
                   variable missing complete     n n_unique                          top_counts ordered
    Front.Facing.Sidewalk.       0    21088 21088        2             Yes: 20309, No: 779, NA: 0   FALSE
                 Graffiti.       0    21088 21088        4   No: 16655, emp: 4214, Yes: 216, Non: 3   FALSE
 Primary.Building.Material       0    21088 21088        6 Bri: 16520, Sid: 3737, Oth: 643, Pla: 116   FALSE
       Structure.Condition   15815     5273 21088        6 NA: 15815, Fai: 2340, Goo: 1523, Poo: 997   FALSE
       Structure.Occupied.       0    21088 21088        8 Occ: 15852, Uno: 4022, Pos: 762, Par: 447   FALSE
             Structure.Use   15787     5301 21088        7  NA: 15787, Res: 4692, Com: 253, Oth: 128   FALSE
         Type.of.Residence   16391     4697 21088        6 NA: 16391, Sin: 3353, Mul: 1090, Apa: 224   FALSE

— Variable type:integer ———————————————————————————————————————————————————————————————————
         variable missing complete     n        mean          sd        p0      p25      p50      p75       p100     hist
 Assessment.Total       0    21088 21088 13467.76    180479.17     -8670     2030     3500     6370    1.2e+07 ▇▁▁▁▁▁▁▁
               Id       0    21088 21088    4.3e+08      1.5e+08   2.1e+07  3.7e+08  4.5e+08  5.3e+08    9.1e+08 ▁▁▁▇▇▁▁▁
  Land.Assessment       0    21088 21088  2421.12     43880.08         0      440      720     1240    3328500 ▇▁▁▁▁▁▁▁
   Land.Use.Value       0    21088 21088  1462.44      1216.37      1000     1110     1110     1120       9112 ▇▁▁▁▁▁▁▁

— Variable type:numeric ———————————————————————————————————————————————————————————————————
                   variable missing complete     n     mean        sd    p0  p25  p50  p75    p100     hist
 Assessment.Improved.Value       0    21088 21088 12042.07 223368.65 -9090 1440 2700 5040  2.1e+07 ▇▁▁▁▁▁▁▁
                  Frontage       0    21088 21088    33.08     18.05     0   25   30   41 458.27   ▇▁▁▁▁▁▁▁
       Number.of.Buildings       0    21088 21088     0.87      9.75  -999    1    1    1  17      ▁▁▁▁▁▁▁▇
              Number.Units       0    21088 21088     1.66      7.94    -2    1    1    2 538      ▇▁▁▁▁▁▁▁
```

2. Data Completeness problems:

Command: skim(df) (see screenshot above)
For missing values, several of the following variables had missing data and had problems with data completeness above the 5% in missing values. Not having these variables would make it hard to make predictions, especially when it comes to pricing. They include:

Structure.Condition: 15,815 missing values (74% missing)



Structure.Use: 15,787 missing values (74% missing)
Type.of.Residence: 16,391 missing values (77% missing)

Command: count(filter(df, Graffiti. == ""))
Graffiti: has 4,214 rows that appear to be just an empty string, but should be considered missing - when looking at summary(df) (19% missing).

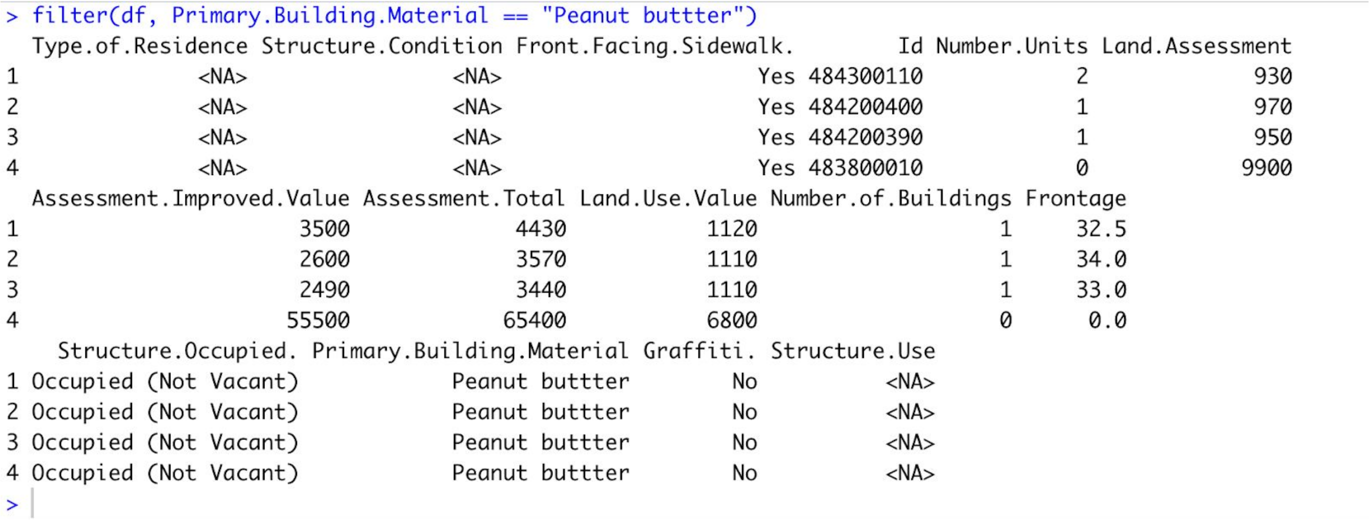
Command: filter(df, Frontage == 0)
Frontage has 2,204 rows with 0 entered (10% of data for Column)

Number.of.Buildings has 1348 rows with 0 (6.87% of data for Column)

While a small %, there were 2 rows with **Number.of.Buildings** with -999.000.

3. Data plausibility problems

Command: summary(df)

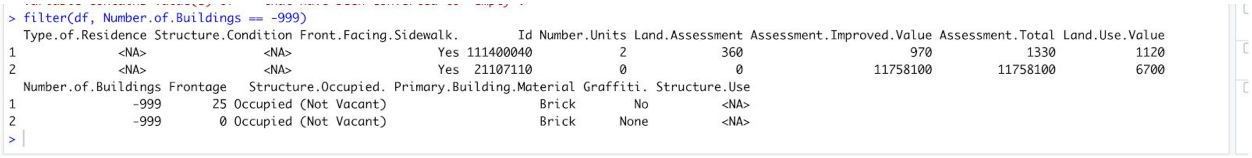


```
> filter(df, Primary.Building.Material == "Peanut butter")
  Type.of.Residence Structure.Condition Front.Facing.Sidewalk.        Id Number.Units Land.Assessment
1              <NA>                <NA>                    Yes 484300110            2             930
2              <NA>                <NA>                    Yes 484200400            1             970
3              <NA>                <NA>                    Yes 484200390            1             950
4              <NA>                <NA>                    Yes 483800010            0            9900
  Assessment.Improved.Value Assessment.Total Land.Use.Value Number.of.Buildings Frontage
1                      3500             4430           1120                   1     32.5
2                      2600             3570           1110                   1     34.0
3                      2490             3440           1110                   1     33.0
4                     55500            65400           6800                   0      0.0
   Structure.Occupied. Primary.Building.Material Graffiti. Structure.Use
1 Occupied (Not Vacant)           Peanut butter        No          <NA>
2 Occupied (Not Vacant)           Peanut butter        No          <NA>
3 Occupied (Not Vacant)           Peanut butter        No          <NA>
4 Occupied (Not Vacant)           Peanut butter        No          <NA>
> 
```

I noticed that **Primary.Building.Material** had 4 rows lising 'Peanut butter' This is not plausible as peanut butter cannot and is not used for constructing buildings or houses.

Command: filter(df, Number.Units == -2.000)
Also, **number.units** has 1 row a value of -2.00

Also, there is an **Assessment.Total** with a value of -8670.

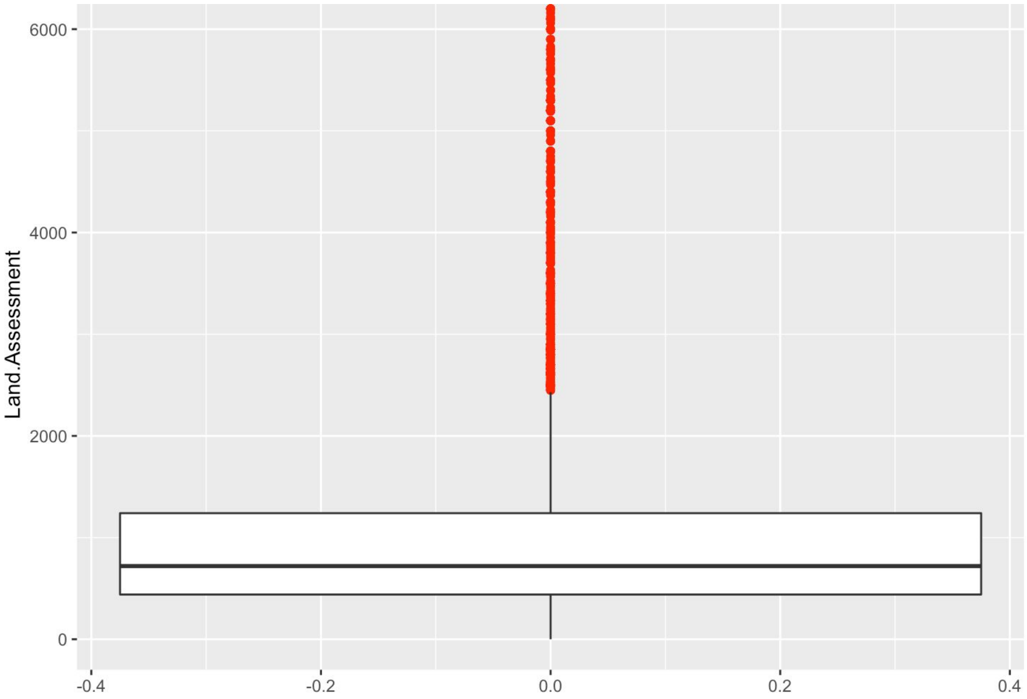


```
> filter(df, Number.of.Buildings == -999)
  Type.of.Residence Structure.Condition Front.Facing.Sidewalk.        Id Number.Units Land.Assessment Assessment.Improved.Value Assessment.Total Land.Use.Value
1              <NA>                <NA>                    Yes 111400040            2             360                       970             1330           1120
2              <NA>                <NA>                    Yes  21107110            0               0                  11758100         11758100           6700
  Number.of.Buildings Frontage   Structure.Occupied. Primary.Building.Material Graffiti. Structure.Use
1                -999       25 Occupied (Not Vacant)                     Brick        No          <NA>
2                -999        0 Occupied (Not Vacant)                     Brick      None          <NA>
> 
```

One of the with Id: 21107110 has an improved value of 11,758,100 and Assessment.Total of 11,758,100 has -999 number of building while having 0 Number.Units.

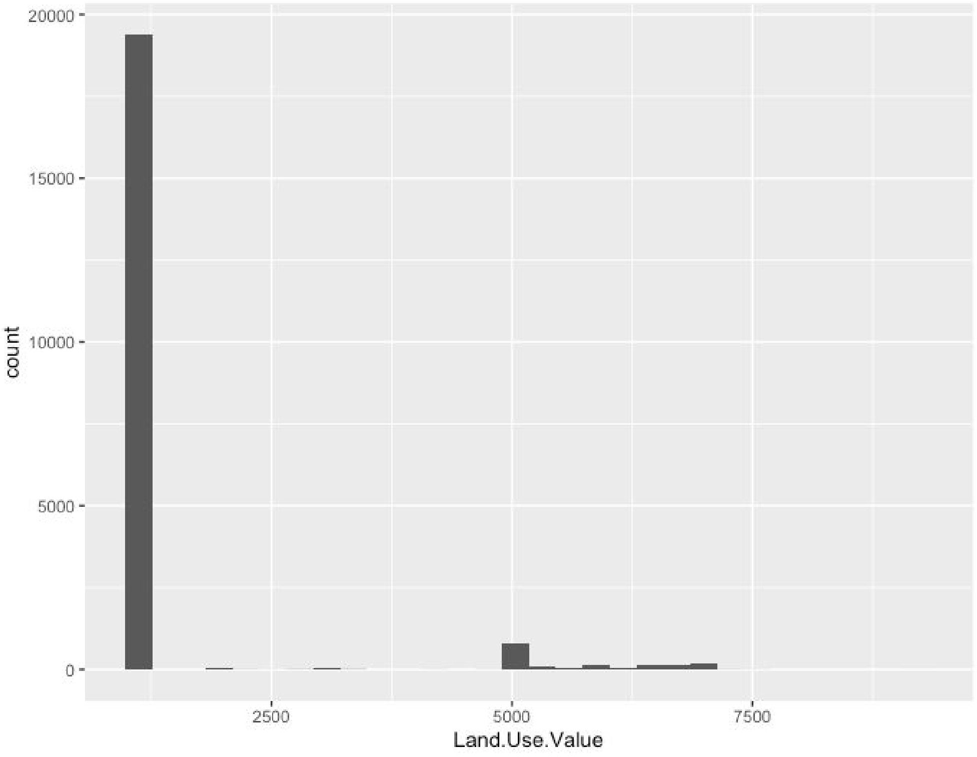
4. **Command to generate box plot:**
ggplot(data = df, aes(y=Land.Assessment)) + geom_boxplot(outlier.color = "red") + coord_cartesian(ylim = c(0,6000))

All points in red are outliers.



5. Land Use value would not be a good candidate for equal width binning.
(9112-1000/4) = 2,028 width
Because Q1 is 1110 Q3 is 1120 and the mean is 1462 mostly all the values would appear to fit in one bin. Therefore, Land Use would not be a good candidate for equal width binning. Equal binning doesn't handle outliers and skewed data well.

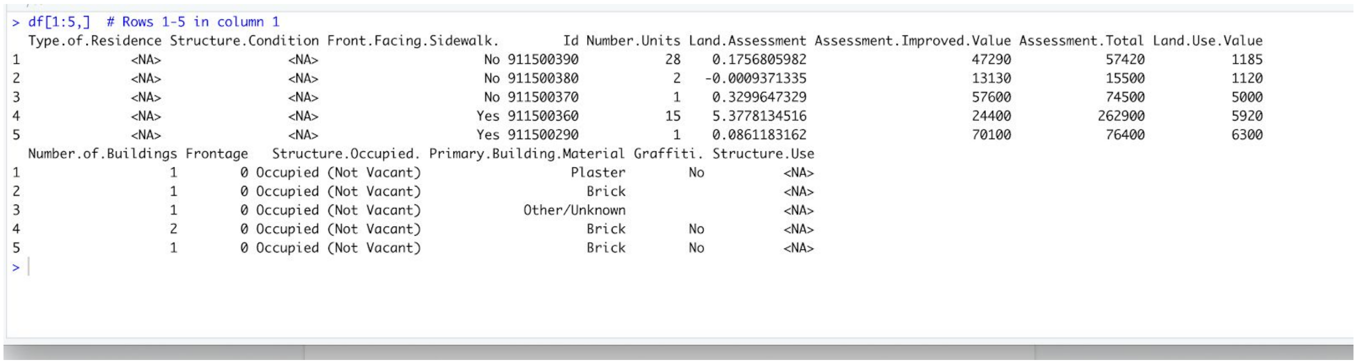
```
ggplot(data = df) + geom_histogram(mapping = aes(Land.Use.Value))
```



6.
Here is the Land Assessment column transformed using z-score normalization. The R **Command used is:**

la <- select(df, Land.Assessment)
laz <- scale(la)
df[,6] <- laz

The first 5 rows are:
Command: df[1:5,]



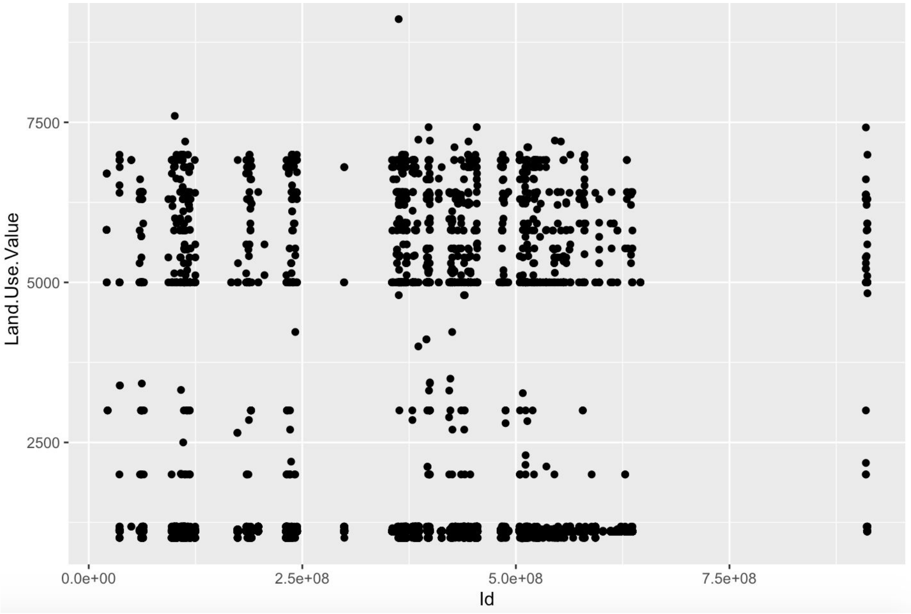
```
> df[1:5,]  # Rows 1-5 in column 1
  Type.of.Residence Structure.Condition Front.Facing.Sidewalk.        Id Number.Units Land.Assessment Assessment.Improved.Value Assessment.Total Land.Use.Value
1             <NA>                <NA>                     No 911500390           28    0.1756805982                     47290            57420           1185
2             <NA>                <NA>                     No 911500380            2   -0.0009371335                     13130            15500           1120
3             <NA>                <NA>                     No 911500370            1    0.3299647329                     57600            74500           5000
4             <NA>                <NA>                    Yes 911500360           15    5.3778134516                     24400           262900           5920
5             <NA>                <NA>                    Yes 911500290            1    0.0861183162                     70100            76400           6300
  Number.of.Buildings Frontage   Structure.Occupied. Primary.Building.Material Graffiti. Structure.Use
1                   1        0 Occupied (Not Vacant)                   Plaster        No          <NA>
2                   1        0 Occupied (Not Vacant)                     Brick                    <NA>
3                   1        0 Occupied (Not Vacant)             Other/Unknown                    <NA>
4                   2        0 Occupied (Not Vacant)                     Brick        No          <NA>
5                   1        0 Occupied (Not Vacant)                     Brick        No          <NA>
> 
```

7. It would **take 4 principal components** to capture 85% of variability in data.
PC1 0.322 + PC2 0.224 + PC3 0.2063 + PC4 0.115 = 0.85 = 85%

Extra credit:

I added many extra ones above. I found this plot of Land Use Value to be interesting.

Land Use Value
ggplot(df, aes(y=Land.Use.Value)) + geom_point()


Land.Use.Value
0.0e+00
2.5e+08
5.0e+08
7.5e+08
Id