data cleaning

Hanyuan

June 14, 2017

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
```

```
#load data
rent <- read.csv('fang88_rent.csv',na.strings=c("NA","NaN", " ","","NULL","N/A"))
house <- read.csv('home.csv',na.strings=c("NA","NaN", " ","", "NULL","N/A"))
```

```
#pick the columns to use in rent and check the statistics of rent dataset
rent_use <- rent[,5:6]
str(rent_use)
```

```
## 'data.frame':    1048575 obs. of  2 variables:
##  $ rentzestimate_amount: int  16619 3495 4448 1430 1400 1375 4556 2500 1850 1650 ...
##  $ unique_id           : Factor w/ 1048571 levels "5.7035E+13","5.71614E+13",..: 614907 614902 75 79
```

```
summary(rent_use)
```

```
##  rentzestimate_amount       unique_id
##  Min.    :   400       5.74132E+13:      2
##  1st Qu.:  1550       5.81157E+13:      2
##  Median :  2200       5.82614E+13:      2
##  Mean   :  2994       5.89261E+13:      2
##  3rd Qu.:  3200       5.7035E+13 :      1
##  Max.   :250000       5.71614E+13:      1
##  NA's   :87721        (Other)    :1048565
```

```
rent_use$unique_id <- as.character(rent_use$unique_id)
```

```
#remove duplicated unique_id
rent_use[duplicated(rent_use$unique_id),]
```

```
##        rentzestimate_amount   unique_id
## 1903                   1800 5.74132E+13
## 8566                     NA 5.81157E+13
## 9170                     NA 5.82614E+13
## 422192                 2962 5.89261E+13
```

```
rent_use <- rent_use %>%
  distinct(unique_id, .keep_all = TRUE)
```

```r
#check the statistics of house dataset
str(house)
```

```
## 'data.frame':    352221 obs. of  13 variables:
##  $ unique_id    : Factor w/ 352221 levels "ABOR_11622550",..: 9 12 33 52 58 59 63 65 73 75 ...
##  $ bathrooms    : num  1 5 3 2 3 2 3 3 3 4 ...
##  $ bedrooms     : int  1 3 3 2 4 3 3 4 3 5 ...
##  $ city         : Factor w/ 5181 levels "29 Palms","3105",..: 191 588 4096 191 191 4175 2483 3559 643
##  $ list_price   : int  225000 6850000 334900 735000 2800000 65000 349990 349990 975000 268004 ...
##  $ latitude     : num  30.4 30.7 29.9 30.3 30.3 ...
##  $ longitude    : num  -98 -98.2 -98 -97.7 -97.8 ...
##  $ property_type: Factor w/ 12 levels "APT","COND","COOP",..: 11 11 2 1 11 11 11 11 11 11 ...
##  $ lot_sqft     : int  0 640 0 0 1 0 NA NA 45 0 ...
##  $ sqft         : int  704 4889 2351 1416 2016 2562 2148 2912 2975 2495 ...
##  $ state        : Factor w/ 54 levels "AK","AL","AR",..: 45 45 45 45 45 45 45 45 45 45 ...
##  $ year_built   : int  1955 2000 2016 2004 1962 1900 2015 2016 1990 2015 ...
##  $ zip          : Factor w/ 7454 levels ".","0","1","1000",..: 5244 5173 5216 5223 5253 5298 5196 521
```

```r
summary(house)
```

```
##          unique_id        bathrooms         bedrooms
##  ABOR_11622550:     1   Min.   : 0.000   Min.   : 0.000
##  ABOR_13245505:     1   1st Qu.: 2.000   1st Qu.: 3.000
##  ABOR_13245677:     1   Median : 2.000   Median : 3.000
##  ABOR_13260708:     1   Mean   : 2.478   Mean   : 3.364
##  ABOR_13811525:     1   3rd Qu.: 3.000   3rd Qu.: 4.000
##  ABOR_14937134:     1   Max.   :99.990   Max.   :99.000
##  (Other)      :352215   NA's   :2651     NA's   :13277
##           city          list_price           latitude
##  Houston    : 15981   Min.   :        0   Min.   :-81.27
##  Miami      :  8342   1st Qu.:   219999   1st Qu.: 29.50
##  Chicago    :  7599   Median :   369900   Median : 33.21
##  Las Vegas  :  4558   Mean   :   690378   Mean   : 33.73
##  Miami Beach:  3412   3rd Qu.:   649900   3rd Qu.: 37.94
##  (Other)    :311840   Max.   :250000000   Max.   : 49.00
##  NA's       :   489   NA's   :1           NA's   :14107
##    longitude        property_type       lot_sqft              sqft
##  Min.   :-159.676   RESI   :248616   Min.   :0.000e+00   Min.   :      0
##  1st Qu.: -97.995   APT    : 56486   1st Qu.:5.600e+03   1st Qu.:   1344
##  Median : -88.117   COND   : 17390   Median :8.712e+03   Median :   2012
##  Mean   : -93.935   RENT   : 10023   Mean   :1.301e+05   Mean   :   2368
##  3rd Qu.: -81.483   LAND   :  9013   3rd Qu.:1.864e+04   3rd Qu.:   2961
##  Max.   :   1.923   (Other): 10692   Max.   :2.147e+09   Max.   :3092760
##  NA's   :14107      NA's   :     1   NA's   :138108      NA's   :25404
##      state         year_built            zip
##  FL     :85162   Min.   :       0   33160  :  2284
##  TX     :81846   1st Qu.:    1970   33139  :  1842
##  CA     :59328   Median :    1993   92253  :  1575
##  IL     :33773   Mean   :    2047   33131  :  1458
##  GA     :30207   3rd Qu.:    2006   77494  :  1181
##  (Other):61234   Max.   :19571939   (Other):343568
##  NA's   :  671   NA's   :16141      NA's   :   313
```

```r
#remove duplicated unique_id
house[duplicated(house$unique_id),]
```

```
##  [1] unique_id     bathrooms     bedrooms      city          list_price
##  [6] latitude      longitude     property_type lot_sqft      sqft
## [11] state         year_built    zip
## <0 rows> (or 0-length row.names)
```

```
house$unique_id <- as.character(house$unique_id)
```

```
#left join house with rent_use by unique_id
house_rent <- merge(house,rent_use, by="unique_id",all.x = TRUE) #left join
str(house_rent)
```

```
## 'data.frame':    352221 obs. of  14 variables:
##  $ unique_id           : chr  "ABOR_11622550" "ABOR_13245505" "ABOR_13245677" "ABOR_13260708" ...
##  $ bathrooms           : num  2 1 1 1 2 3 3 5 1 2 ...
##  $ bedrooms            : int  2 1 1 1 2 4 4 4 1 2 ...
##  $ city                : Factor w/ 5181 levels "29 Palms","3105",..: 4662 1716 1716 1716 550 2483 19
##  $ list_price          : int  249900 79500 79500 119000 342850 700000 550000 1899000 225000 995000 .
##  $ latitude            : num  30.9 30.6 30.6 30.6 30.7 ...
##  $ longitude           : num  -98.5 -98.4 -98.4 -98.4 -98.4 ...
##  $ property_type       : Factor w/ 12 levels "APT","COND","COOP",..: 11 1 1 1 11 11 11 11 11 11 ...
##  $ lot_sqft            : int  1 NA NA NA 1 2 4 102 0 77 ...
##  $ sqft                : int  2324 957 957 957 1592 2808 2500 5692 704 1950 ...
##  $ state               : Factor w/ 54 levels "AK","AL","AR",..: 45 45 45 45 45 45 45 45 45 45 ...
##  $ year_built          : int  1950 1982 1982 1982 2000 1993 2015 2002 1955 1984 ...
##  $ zip                 : Factor w/ 7454 levels ".","0","1","1000",..: 5219 5205 5205 5205 5171 5196 5
##  $ rentzestimate_amount: int  NA 950 900 950 1787 3087 NA 10632 1350 5416 ...
```

```
summary(house_rent)
```

```
##   unique_id           bathrooms         bedrooms              city
##  Length:352221      Min.   : 0.000   Min.   : 0.000   Houston    : 15981
##  Class :character   1st Qu.: 2.000   1st Qu.: 3.000   Miami      :  8342
##  Mode  :character   Median : 2.000   Median : 3.000   Chicago    :  7599
##                     Mean   : 2.478   Mean   : 3.364   Las Vegas  :  4558
##                     3rd Qu.: 3.000   3rd Qu.: 4.000   Miami Beach:  3412
##                     Max.   :99.990   Max.   :99.000   (Other)    :311840
##                     NA's   :2651     NA's   :13277    NA's       :   489
##    list_price           latitude       longitude        property_type
##  Min.   :        0   Min.   :-81.27   Min.   :-159.676   RESI   :248616
##  1st Qu.:   219999   1st Qu.: 29.50   1st Qu.: -97.995   APT    : 56486
##  Median :   369900   Median : 33.21   Median : -88.117   COND   : 17390
##  Mean   :   690378   Mean   : 33.73   Mean   : -93.935   RENT   : 10023
##  3rd Qu.:   649900   3rd Qu.: 37.94   3rd Qu.: -81.483   LAND   :  9013
##  Max.   :250000000   Max.   : 49.00   Max.   :   1.923   (Other): 10692
##  NA's   :1           NA's   :14107    NA's   :14107      NA's   :     1
##     lot_sqft              sqft             state         year_built
##  Min.   :0.000e+00   Min.   :      0   FL     :85162   Min.   :       0
##  1st Qu.:5.600e+03   1st Qu.:   1344   TX     :81846   1st Qu.:    1970
##  Median :8.712e+03   Median :   2012   CA     :59328   Median :    1993
##  Mean   :1.301e+05   Mean   :   2368   IL     :33773   Mean   :    2047
##  3rd Qu.:1.864e+04   3rd Qu.:   2961   GA     :30207   3rd Qu.:    2006
##  Max.   :2.147e+09   Max.   :3092760   (Other):61234   Max.   :19571939
##  NA's   :138108      NA's   :25404     NA's   :  671   NA's   :16141
##       zip         rentzestimate_amount
##  33160  :  2284   Min.   :   400
##  33139  :  1842   1st Qu.:  1600
```

```
##  92253   :  1575   Median :  2300
##  33131   :  1458   Mean   :  3854
##  77494   :  1181   3rd Qu.:  3500
##  (Other):343568   Max.   :250000
##  NA's    :   313   NA's   :164944
```

```
#create a function imp.median to replace NA with median
imp.median <- function (a){
  missing <- is.na(a)
  imputed <- a
  imputed[missing] <- median(a,  na.rm=TRUE)
  return (imputed)
}
```

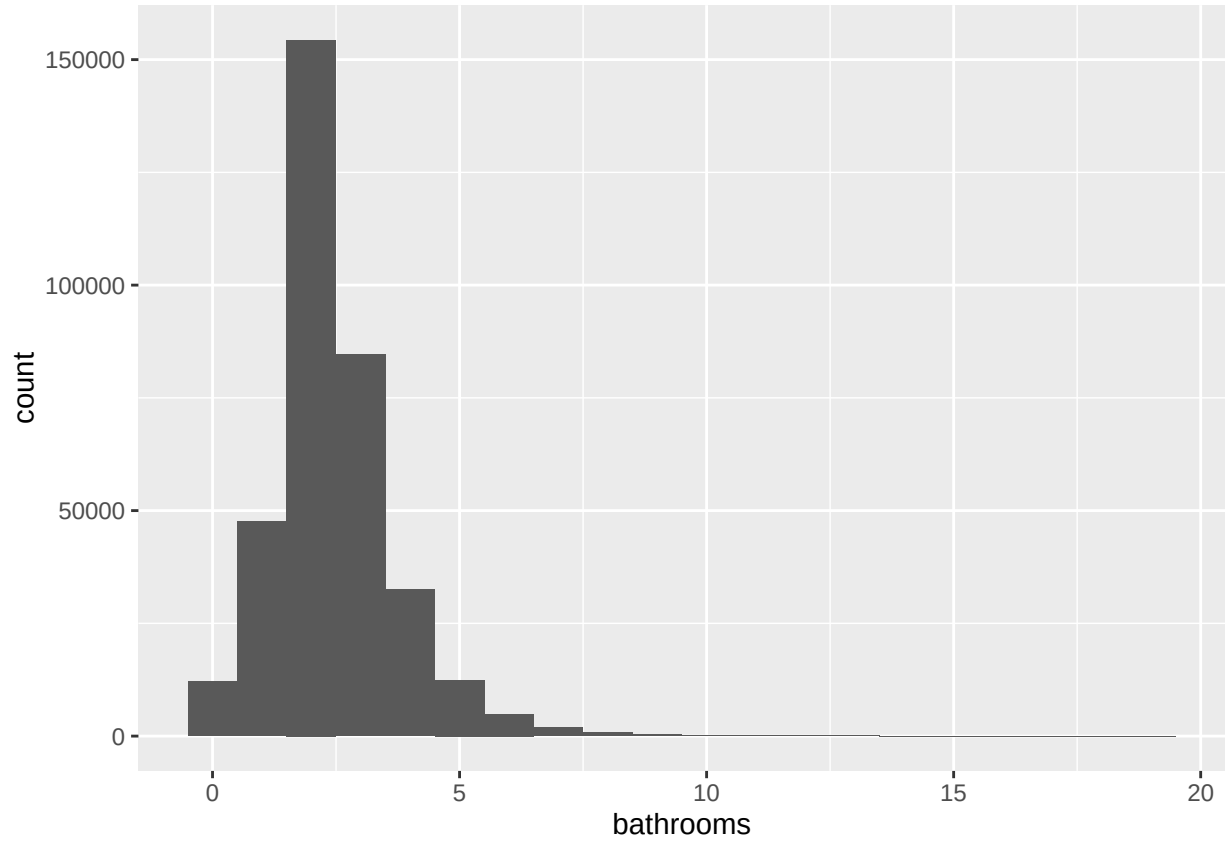
```
#cleaning of feature:bathrooms
house_rent$bathrooms <- imp.median(house_rent$bathrooms) # interpolate median

quantile(house_rent$bathrooms, c(0,0.01,0.9999,1), na.rm = TRUE) #capping of outliers--0.9999
```

```
##     0%     1% 99.99%   100%
##   0.00   0.00  19.00  99.99
```

```
house_rent$bathrooms[house_rent$bathrooms > quantile(house_rent$bathrooms,0.9999, na.rm = TRUE)] =
  quantile(house_rent$bathrooms, 0.9999, na.rm = TRUE)

# visulization
ggplot(house_rent, aes(bathrooms)) +
  geom_histogram(binwidth = 1)
```



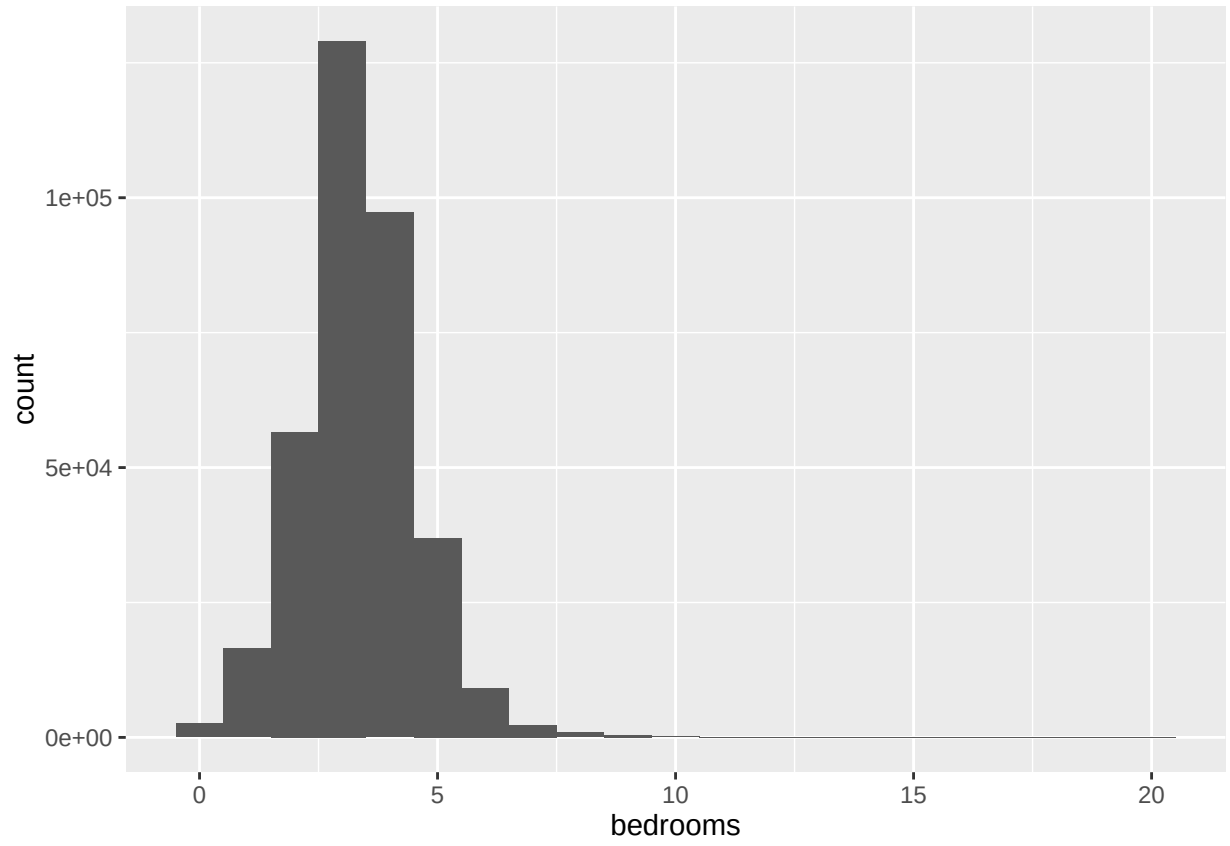
```r
#cleaning of feature:bedrooms
house_rent$bedrooms <- imp.median(house_rent$bedrooms) #interpolate median

quantile(house_rent$bedrooms, c(0, 0.01, 0.9999,1), na.rm = TRUE) #capping of outliers--0.9999
```

```
##     0%     1% 99.99%   100%
##      0      1     20     99
```

```r
house_rent$bedrooms[house_rent$bedrooms > quantile(house_rent$bedrooms,0.9999, na.rm = TRUE)] =
  quantile(house_rent$bedrooms, 0.9999, na.rm = TRUE)

#visualization
ggplot(house_rent, aes(bedrooms)) +
  geom_histogram(binwidth = 1)
```



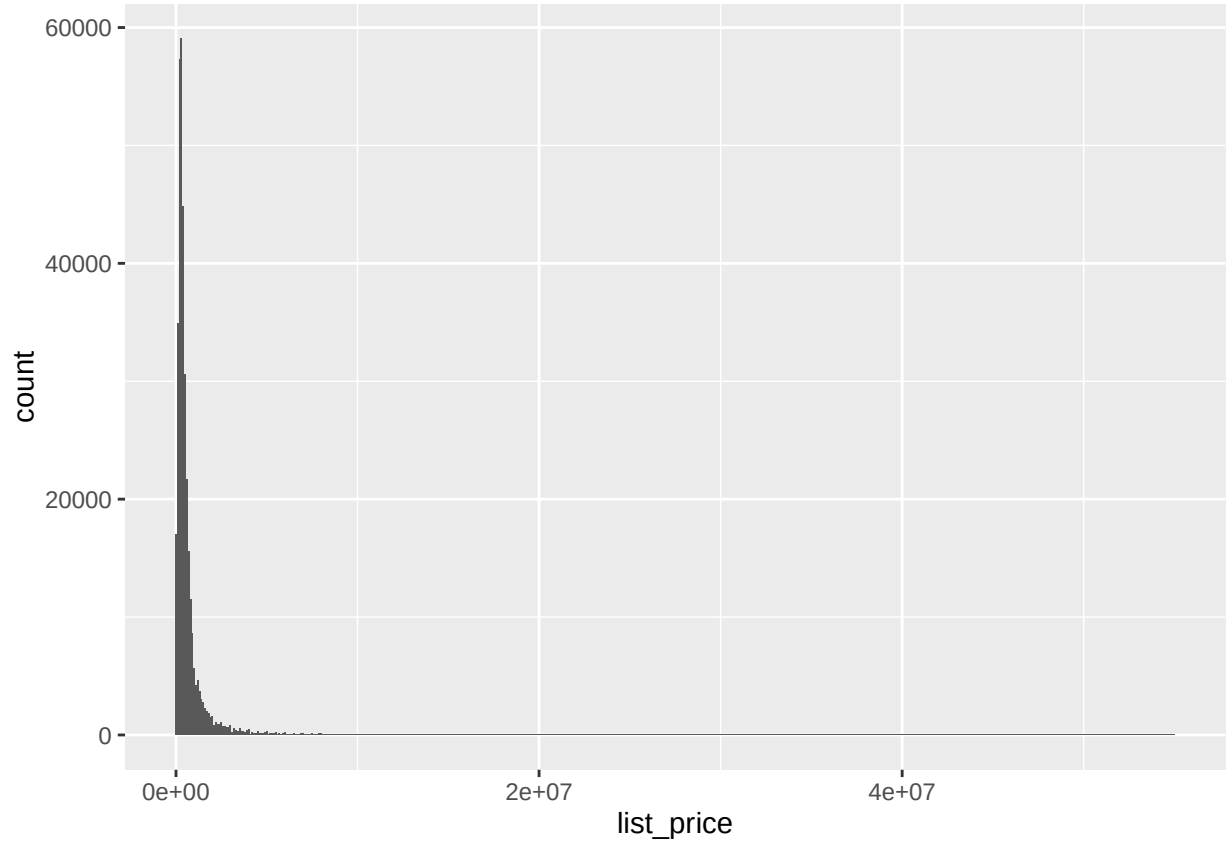

```r
#cleaning of feature:list_price
house_rent$list_price <- imp.median(house_rent$list_price) #interpolate median

quantile(house_rent$list_price, c(0,0.01,0.9999,1), na.rm = TRUE) #capping of outliers--0.9999
```

```
##       0%       1%   99.99%     100%
## 0.00e+00 1.45e+03 5.50e+07 2.50e+08
```

```r
house_rent$list_price[house_rent$list_price > quantile(house_rent$list_price,0.9999, na.rm = TRUE)] =
  quantile(house_rent$list_price, 0.9999, na.rm = TRUE)
#visualization -- right skewed
ggplot(house_rent, aes(list_price)) +
```

```
  geom_histogram(binwidth = 100000)
```




```
#cleaning of feature:latitude and longtitude
house_rent$latitude <- imp.median(house_rent$latitude) #interpolate median
house_rent$longitude <- imp.median(house_rent$longitude)
```

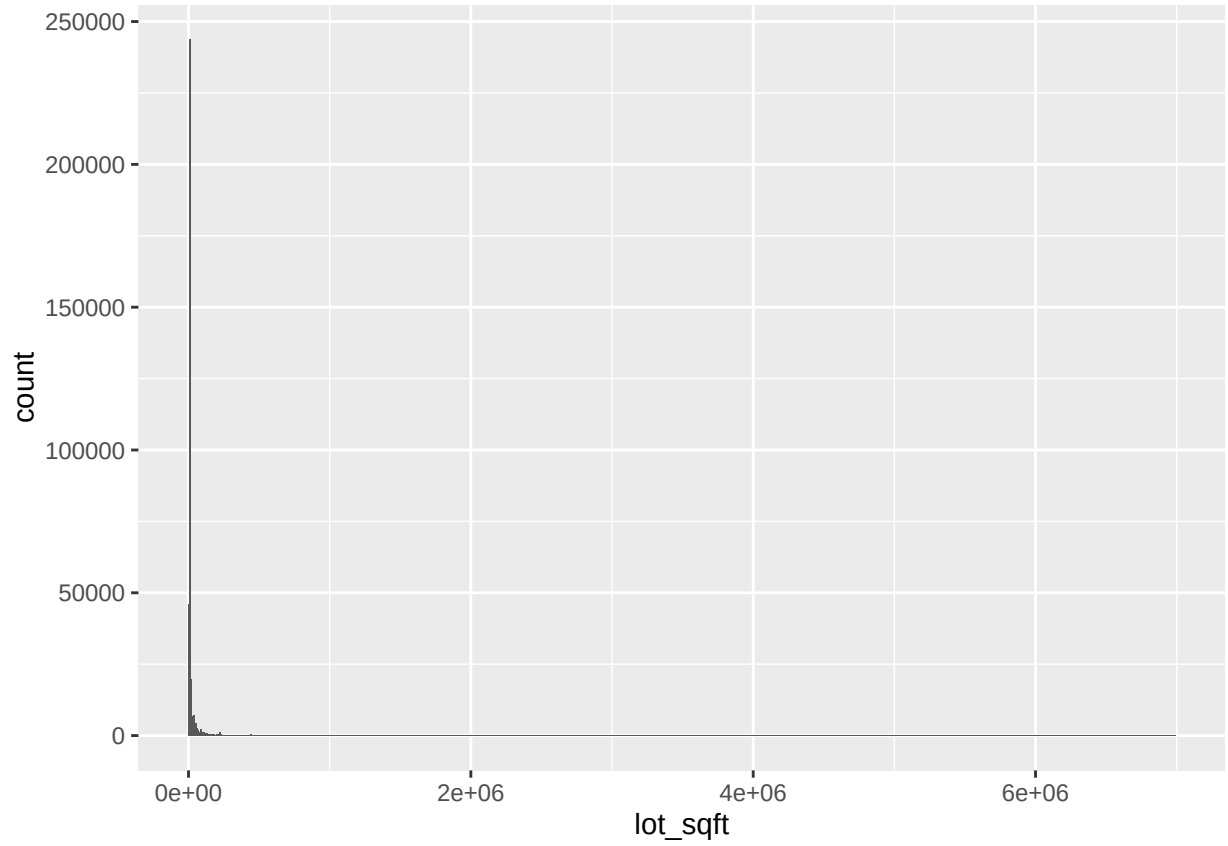
```
#cleaning of feature:lot_sqft
house_rent$lot_sqft <- imp.median(house_rent$lot_sqft)#interpolate median

quantile(house_rent$lot_sqft, c(0,0.01,0.9995,1), na.rm = TRUE)#capping of outliers--0.9995
```

```
##         0%         1%     99.95%       100%
##          0          0    6992861 2147483647
```

```
house_rent$lot_sqft[house_rent$lot_sqft > quantile(house_rent$lot_sqft,0.9995, na.rm = TRUE)] =
  quantile(house_rent$lot_sqft, 0.9995, na.rm = TRUE)

# visualization -- right-skewed
ggplot(house_rent, aes(lot_sqft)) +
  geom_histogram(binwidth = 10000)
```



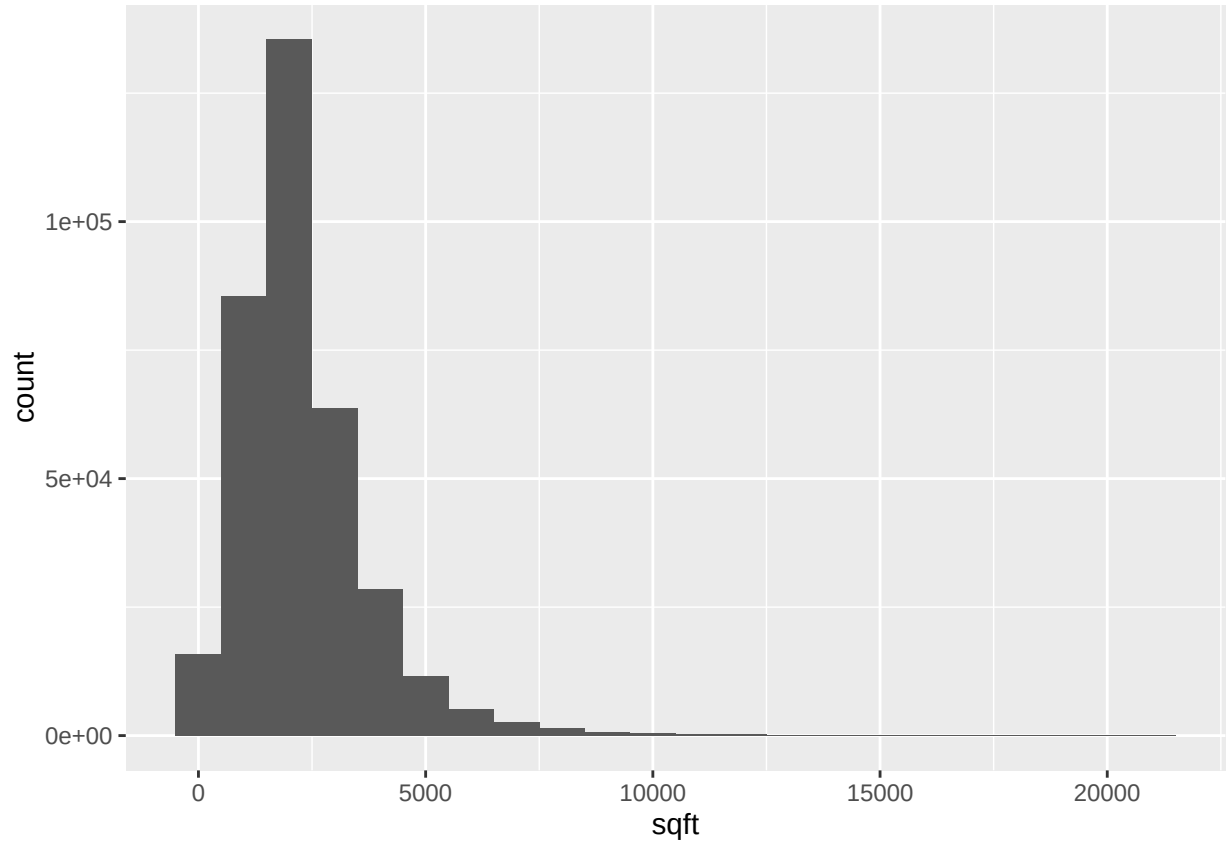

```
#cleaning of feature:sqft
house_rent$sqft <- imp.median(house_rent$sqft) #interpolate median

quantile(house_rent$sqft, c(0,0.01,0.9995,1), na.rm = TRUE) #capping of outliers--0.9995
```

```
##          0%          1%      99.95%        100%
##        0.00        0.00    20687.21 3092760.00
```

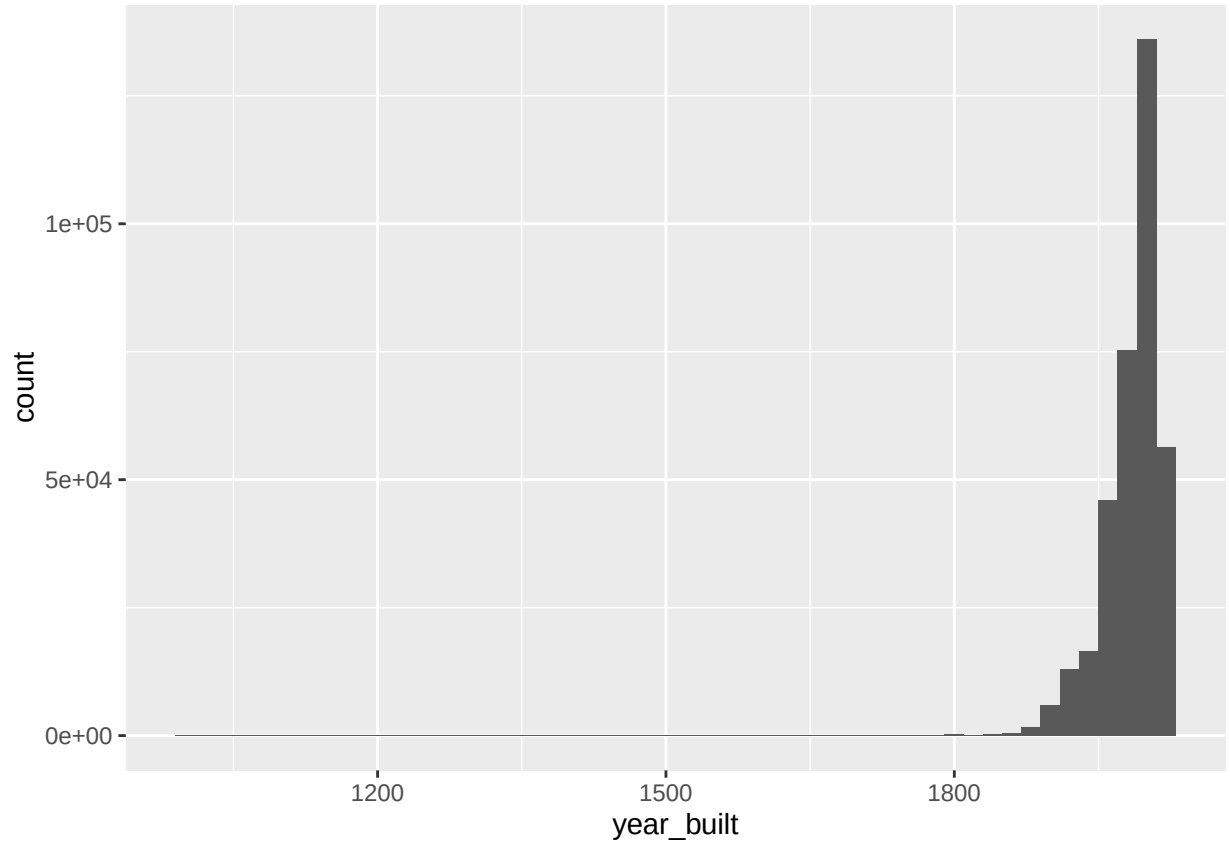
```
house_rent$sqft[house_rent$sqft > quantile(house_rent$sqft,0.9995, na.rm = TRUE)] =
  quantile(house_rent$sqft, 0.9995, na.rm = TRUE)

#visualization -- right-skewed
ggplot(house_rent, aes(sqft)) +
  geom_histogram(binwidth = 1000)
```




```
#cleaning of feature:year_built
house_rent$year_built <- imp.median(house_rent$year_built) #interpolate median

#for year_built, earlier than 1000 and later than 2017 will be replaced by median
house_rent$year_built[house_rent$year_built < 1000 | house_rent$year_built > 2017] <-
  median(house_rent$year_built)
#visualization -- left-skewed
ggplot(house_rent, aes(year_built)) +
  geom_histogram(binwidth = 20)
```




```
#cleaning of feature: state and zip code
#chose rows with state/zip not being NA
#zip code completion for zip codes equals 5 digits (add 0 in the front)
length(house_rent[as.numeric(as.character(house_rent$zip))<=1000,]$zip)
```

```
## Warning in `[.data.frame`(house_rent, as.numeric(as.character(house_rent
## $zip)) <= : NAs introduced by coercion
```

```
## [1] 622
```

```
#since those less than 4 digits are of small number, so delete the records

house_rent$state[house_rent$state == 'hi'] = "HI"

house_rent <- subset(house_rent, state %in% c( "AK","AL","AR","AZ","CA","CO","CT","DC","DE","FL","GA","I
                                               "IL","IN","KS","KY","LA","MA","MD","ME","MI","MN","MO","MS
                                               "ND","NE","NH","NJ","NM","NV","NY","OH","OK","OR","PA","RI
                                               "TN","TX","UT","VA", "VT","WA", "WI", "WV", "WY",
                                               "XX","Hi","Unk","Ha","BJ"))


house_rent <- house_rent[as.numeric(as.character(house_rent$zip))>=10000 &
                           as.numeric(as.character(house_rent$zip))<=99999,]
```

```
## Warning in `[.data.frame`(house_rent, as.numeric(as.character(house_rent
## $zip)) >= : NAs introduced by coercion
```

```
## Warning in `[.data.frame`(house_rent, as.numeric(as.character(house_rent
```

```
## $zip)) >= : NAs introduced by coercion
```

```
house_rent<- house_rent[!is.na(house_rent$zip),]

house_rent<- house_rent[!is.na(house_rent$property_type),]
house_rent$property_type[house_rent$property_type == 'Other'] = "OTHER"
```

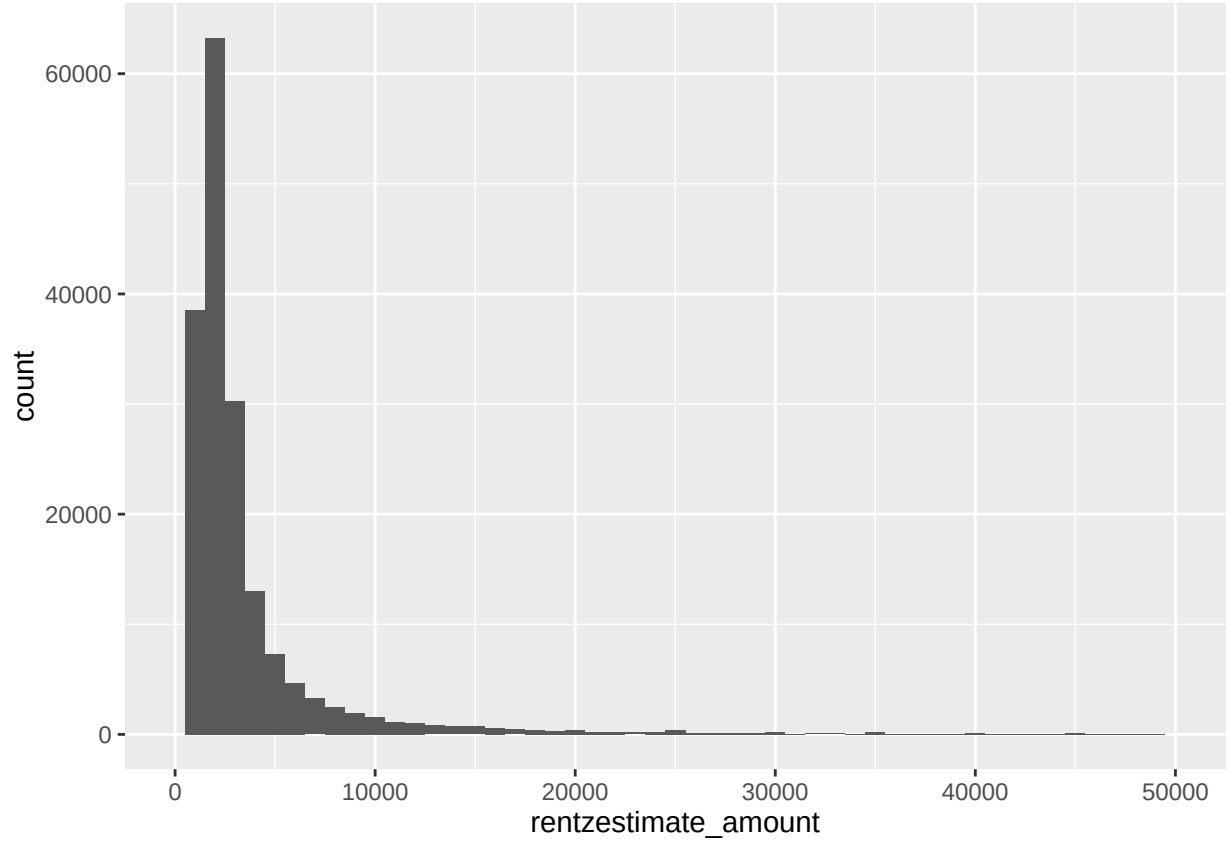
```
#cleaning of feature:rentzestimate_amount
#house_rent$rentzestimate_amount <- imp.median(house_rent$rentzestimate_amount) #interpolate median

quantile(house_rent$rentzestimate, c(0,0.01,0.9999,1), na.rm = TRUE)
```

```
##       0%       1%   99.99%     100%
##    400.0    850.0 169176.7 250000.0
```

```
#visualization -- right-skewed
ggplot(house_rent, aes(rentzestimate_amount)) +
  geom_histogram(binwidth = 1000) +
  xlim(0,50000)
```

```
## Warning: Removed 153537 rows containing non-finite values (stat_bin).
```




```
#checked the statistics of the cleaned dataset again
summary(house_rent)
```

```
##   unique_id           bathrooms         bedrooms                city
##  Length:329503      Min.   : 0.000   Min.   : 0.000   Houston    : 15981
##  Class :character   1st Qu.: 2.000   1st Qu.: 3.000   Miami      :  8341
##  Mode  :character   Median : 2.000   Median : 3.000   Chicago    :  7599
```

```
##                    Mean   : 2.486   Mean   : 3.341   Las Vegas  :  4558
##                    3rd Qu.: 3.000   3rd Qu.: 4.000   Miami Beach:  3410
##                    Max.   :19.000   Max.   :20.000   (Other)    :289349
##                                                      NA's       :   265
##    list_price          latitude        longitude        property_type
##  Min.   :       1   Min.   :-81.27   Min.   :-159.676   RESI   :233188
##  1st Qu.:  216900   1st Qu.: 29.34   1st Qu.: -98.817   APT    : 51846
##  Median :  365000   Median : 33.09   Median : -88.274   COND   : 16907
##  Mean   :  685455   Mean   : 33.18   Mean   : -95.146   RENT   :  9920
##  3rd Qu.:  649000   3rd Qu.: 34.44   3rd Qu.: -82.402   LAND   :  8975
##  Max.   :55000000   Max.   : 49.00   Max.   :   1.923   MULT   :  5681
##                                                        (Other):  2986
##     lot_sqft            sqft           state         year_built
##  Min.   :      0   Min.   :    0   FL     :85012   Min.   :1000
##  1st Qu.:   7405   1st Qu.: 1391   TX     :81833   1st Qu.:1973
##  Median :   8712   Median : 2012   CA     :59318   Median :1993
##  Mean   :  32661   Mean   : 2297   IL     :33773   Mean   :1987
##  3rd Qu.:  10000   3rd Qu.: 2863   GA     :30207   3rd Qu.:2006
##  Max.   :6992861   Max.   :20687   WA     :14658   Max.   :2017
##                                    (Other):24702
##       zip         rentzestimate_amount
##  33160  :  2284   Min.   :   400
##  33139  :  1842   1st Qu.:  1600
##  92253  :  1575   Median :  2300
##  33131  :  1458   Mean   :  3884
##  77494  :  1181   3rd Qu.:  3520
##  33180  :  1175   Max.   :250000
##  (Other):319988   NA's   :153083
```

```
#save as csv
write.csv(house_rent,'house_rent_2.csv')
```