

# HW1 - Assignment 1

Naveen Nagarajan

2/14/2021

## Problem 1

Explore realldirect.com thinking about how buyers and sellers would navigate, and how the website is organized. Use the datasets provided for Bronx, Brooklyn, Manhattan, Queens, and Staten Island. Do the following: • Load in and clean the data. • Conduct exploratory data analysis in order to find out where there are outliers or missing values, decide how you will treat them, make sure the dates are formatted correctly, make sure values you think are numerical are being treated as such, etc. • Conduct exploratory data analysis to visualize and make comparisons for residential building category classes across boroughs and across time (select the following: 1-, 2-, and 3-family homes, coops, and condos). Use histograms, boxplots, scatterplots or other visual graphs. Provide summary statistics along with your conclusions.

### Load and clean data sets

#### Total records loaded for individual boroughs

Summary of stats on the data frame displays class of sale price, gross. Sq. ft, land. Sq. Ft and sale date is character. Transformation needs to be applied to convert these to numeric and date fields respectively

```
summary(bdset)
```

```
##      borough      neighborhood      building.class.category
## Length:85975      Length:85975      Length:85975
## Class :character  Class :character  Class :character
## Mode  :character  Mode  :character  Mode  :character
##
##
##
## tax.class.at.present      block      lot      ease.ment
## Length:85975      Min. : 1      Min. : 1.0      Length:85975
## Class :character  1st Qu.: 1052      1st Qu.: 23.0      Class :character
## Mode  :character  Median : 2157      Median : 51.0      Mode  :character
##                      Mean : 3661      Mean : 405.4
##                      3rd Qu.: 5599      3rd Qu.:1009.0
##                      Max. :16323      Max. :9117.0
##
## building.class.at.present      address      apart.ment.number
## Length:85975      Length:85975      Length:85975
## Class :character      Class :character  Class :character
## Mode  :character      Mode  :character  Mode  :character
```

```
##
##
##
##
##      zip.code      residential.units commercial.units      total.units
## Min.       :    0   Min.       : 0.00   Min.       : 0.0000   Min.       : 0.000
## 1st Qu.:10028   1st Qu.: 0.00   1st Qu.: 0.0000   1st Qu.: 1.000
## Median :11201   Median : 1.00   Median : 0.0000   Median : 1.000
## Mean      :10758   Mean      : 1.96   Mean      : 0.2349   Mean      : 2.252
## 3rd Qu.:11238   3rd Qu.: 2.00   3rd Qu.: 0.0000   3rd Qu.: 2.000
## Max.      :11694   Max.      :904.00   Max.      :604.0000   Max.      :904.000
##                NA's      :3          NA's      :1          NA's      :4
## land.square.feet  gross.square.feet    year.built  tax.class.at.time.of.sale
## Length:85975      Length:85975          Min.       : 0   Min.       :1.000
## Class :character   Class :character    1st Qu.:1910   1st Qu.:1.000
## Mode  :character   Mode  :character    Median :1931   Median :2.000
##                                     Mean      :1681   Mean      :1.868
##                                     3rd Qu.:1964   3rd Qu.:2.000
##                                     Max.      :2013   Max.      :4.000
##                                     NA's      :1
## building.class.at.time.of.sale  sale.price      sale.date
## Length:85975                    Length:85975      Length:85975
## Class :character                 Class :character   Class :character
## Mode  :character                 Mode  :character   Mode  :character
##
##
##
##
```

Define types and format date

```
# define types - numeric, date
bdset$land.square.feet <- as.numeric(gsub("[^[:digit:]]", "", bdset$land.square.feet))
bdset$gross.square.feet <- as.numeric(bdset$gross.square.feet)
```

## Warning: NAs introduced by coercion

```
bdset$sale.price <- as.numeric(gsub("[^[:digit:]]", "", bdset$sale.price))
bdset$sale.date <- as.Date(bdset$sale.date)
bdset$building.class.category <- trim(bdset$building.class.category)

summary(bdset)
```

```
##      borough      neighborhood      building.class.category
## Length:85975      Length:85975      Length:85975
## Class :character   Class :character   Class :character
## Mode  :character   Mode  :character   Mode  :character
##
##
##
##
## tax.class.at.present      block      lot      ease.ment
## Length:85975              Min.       : 1   Min.       : 1.0   Length:85975
```

```
## Class :character      1st Qu.: 1052   1st Qu.: 23.0   Class :character
## Mode :character      Median : 2157   Median : 51.0   Mode :character
##                               Mean : 3661   Mean : 405.4
##                               3rd Qu.: 5599   3rd Qu.:1009.0
##                               Max. :16323   Max. :9117.0
##
## building.class.at.present address      apart.ment.number
## Length:85975          Length:85975      Length:85975
## Class :character      Class :character   Class :character
## Mode :character      Mode :character    Mode :character
##
##
##
## zip.code      residential.units commercial.units      total.units
## Min. : 0      Min. : 0.00      Min. : 0.0000      Min. : 0.000
## 1st Qu.:10028  1st Qu.: 0.00      1st Qu.: 0.0000      1st Qu.: 1.000
## Median :11201  Median : 1.00      Median : 0.0000      Median : 1.000
## Mean :10758   Mean : 1.96      Mean : 0.2349      Mean : 2.252
## 3rd Qu.:11238  3rd Qu.: 2.00      3rd Qu.: 0.0000      3rd Qu.: 2.000
## Max. :11694   Max. :904.00      Max. :604.0000      Max. :904.000
##                               NA's :3      NA's :1      NA's :4
## land.square.feet gross.square.feet year.built      tax.class.at.time.of.sale
## Min. : 0      Min. : 0.00      Min. : 0      Min. :1.000
## 1st Qu.: 0      1st Qu.: 0.00      1st Qu.:1910      1st Qu.:1.000
## Median : 1512   Median : 0.00      Median :1931      Median :2.000
## Mean : 3085   Mean : 26.17      Mean :1681      Mean :1.868
## 3rd Qu.: 2625   3rd Qu.: 0.00      3rd Qu.:1964      3rd Qu.:2.000
## Max. :7446955   Max. :999.00      Max. :2013      Max. :4.000
##                               NA's :41789      NA's :1
## building.class.at.time.of.sale sale.price      sale.date
## Length:85975          Min. : 0      Min. :2012-08-01
## Class :character      1st Qu.: 0      1st Qu.:2012-11-07
## Mode :character      Median : 260000 Median :2013-01-24
##                               Mean : 885098 Mean :2013-01-30
##                               3rd Qu.: 620000 3rd Qu.:2013-05-02
##                               Max. :1307965050 Max. :2013-08-26
##
```

Missing values

```
# missing values
unique(bdset$building.class.category)
```

```
## [1] "01 ONE FAMILY HOMES"
## [2] "02 TWO FAMILY HOMES"
## [3] "03 THREE FAMILY HOMES"
## [4] "05 TAX CLASS 1 VACANT LAND"
## [5] "07 RENTALS - WALKUP APARTMENTS"
## [6] "10 COOPS - ELEVATOR APARTMENTS"
## [7] "14 RENTALS - 4-10 UNIT"
## [8] "22 STORE BUILDINGS"
## [9] "27 FACTORIES"
```

```
## [10] "29  COMMERCIAL GARAGES"
## [11] "30  WAREHOUSES"
## [12] "31  COMMERCIAL VACANT LAND"
## [13] "41  TAX CLASS 4 - OTHER"
## [14] "04  TAX CLASS 1 CONDOS"
## [15] "06  TAX CLASS 1 - OTHER"
## [16] "21  OFFICE BUILDINGS"
## [17] "28  COMMERCIAL CONDOS"
## [18] "08  RENTALS - ELEVATOR APARTMENTS"
## [19] "09  COOPS - WALKUP APARTMENTS"
## [20] "13  CONDOS - ELEVATOR APARTMENTS"
## [21] ""
## [22] "12  CONDOS - WALKUP APARTMENTS"
## [23] "36  OUTDOOR RECREATIONAL FACILITIES"
## [24] "11A CONDO-RENTALS"
## [25] "38  ASYLUMS AND HOMES"
## [26] "37  RELIGIOUS FACILITIES"
## [27] "17  CONDOPS"
## [28] "32  HOSPITAL AND HEALTH FACILITIES"
## [29] "40  SELECTED GOVERNMENTAL FACILITIES"
## [30] "33  EDUCATIONAL FACILITIES"
## [31] "35  INDOOR PUBLIC AND CULTURAL FACILITIES"
## [32] "15  CONDOS - 2-10 UNIT RESIDENTIAL"
## [33] "18  TAX CLASS 3 - UTILITY PROPERTIES"
## [34] "16  CONDOS - 2-10 UNIT WITH COMMERCIAL UNIT"
## [35] "23  LOFT BUILDINGS"
## [36] "26  OTHER HOTELS"
## [37] "25  LUXURY HOTELS"
## [38] "34  THEATRES"
## [39] "24  TAX CLASS 4 - UTILITY BUREAU PROPERTIES"
## [40] "11  SPECIAL CONDO BILLING LOTS"
```

```
bdset <- bdset %>% mutate(building.class.category = na_if(building.class.category,
  ""))
unique(bdset$building.class.category)
```

```
## [1] "01  ONE FAMILY HOMES"
## [2] "02  TWO FAMILY HOMES"
## [3] "03  THREE FAMILY HOMES"
## [4] "05  TAX CLASS 1 VACANT LAND"
## [5] "07  RENTALS - WALKUP APARTMENTS"
## [6] "10  COOPS - ELEVATOR APARTMENTS"
## [7] "14  RENTALS - 4-10 UNIT"
## [8] "22  STORE BUILDINGS"
## [9] "27  FACTORIES"
## [10] "29  COMMERCIAL GARAGES"
## [11] "30  WAREHOUSES"
## [12] "31  COMMERCIAL VACANT LAND"
## [13] "41  TAX CLASS 4 - OTHER"
## [14] "04  TAX CLASS 1 CONDOS"
## [15] "06  TAX CLASS 1 - OTHER"
## [16] "21  OFFICE BUILDINGS"
## [17] "28  COMMERCIAL CONDOS"
## [18] "08  RENTALS - ELEVATOR APARTMENTS"
```

```
## [19] "09 COOPS - WALKUP APARTMENTS"
## [20] "13 CONDOS - ELEVATOR APARTMENTS"
## [21] NA
## [22] "12 CONDOS - WALKUP APARTMENTS"
## [23] "36 OUTDOOR RECREATIONAL FACILITIES"
## [24] "11A CONDO-RENTALS"
## [25] "38 ASYLUMS AND HOMES"
## [26] "37 RELIGIOUS FACILITIES"
## [27] "17 CONDOPS"
## [28] "32 HOSPITAL AND HEALTH FACILITIES"
## [29] "40 SELECTED GOVERNMENTAL FACILITIES"
## [30] "33 EDUCATIONAL FACILITIES"
## [31] "35 INDOOR PUBLIC AND CULTURAL FACILITIES"
## [32] "15 CONDOS - 2-10 UNIT RESIDENTIAL"
## [33] "18 TAX CLASS 3 - UTILITY PROPERTIES"
## [34] "16 CONDOS - 2-10 UNIT WITH COMMERCIAL UNIT"
## [35] "23 LOFT BUILDINGS"
## [36] "26 OTHER HOTELS"
## [37] "25 LUXURY HOTELS"
## [38] "34 THEATRES"
## [39] "24 TAX CLASS 4 - UTILITY BUREAU PROPERTIES"
## [40] "11 SPECIAL CONDO BILLING LOTS"
```

```
bdset$building.class.category <- as.factor(bdset$building.class.category)
summary(bdset$building.class.category)
```

```
##              01  ONE FAMILY HOMES
##              14846
##              02  TWO FAMILY HOMES
##              13678
##              03  THREE FAMILY HOMES
##              4135
##              04  TAX CLASS 1 CONDOS
##              1251
##              05  TAX CLASS 1 VACANT LAND
##              1230
##              06  TAX CLASS 1 - OTHER
##              180
##              07  RENTALS - WALKUP APARTMENTS
##              3989
##              08  RENTALS - ELEVATOR APARTMENTS
##              581
##              09  COOPS - WALKUP APARTMENTS
##              2600
##              10  COOPS - ELEVATOR APARTMENTS
##              13771
##              11  SPECIAL CONDO BILLING LOTS
##              1
##              11A CONDO-RENTALS
##              34
##              12  CONDOS - WALKUP APARTMENTS
##              929
##              13  CONDOS - ELEVATOR APARTMENTS
##              13313
```

```

##          14 RENTALS - 4-10 UNIT
##                      821
##          15 CONDOS - 2-10 UNIT RESIDENTIAL
##                      1123
## 16 CONDOS - 2-10 UNIT WITH COMMERCIAL UNIT
##                      85
##                      17 CONDOPS
##                      1415
##          18 TAX CLASS 3 - UTILITY PROPERTIES
##                      6
##          21 OFFICE BUILDINGS
##                      432
##          22 STORE BUILDINGS
##                      1244
##          23 LOFT BUILDINGS
##                      144
## 24 TAX CLASS 4 - UTILITY BUREAU PROPERTIES
##                      1
##          25 LUXURY HOTELS
##                      1647
##          26 OTHER HOTELS
##                      40
##          27 FACTORIES
##                      346
##          28 COMMERCIAL CONDOS
##                      1682
##          29 COMMERCIAL GARAGES
##                      888
##          30 WAREHOUSES
##                      430
##          31 COMMERCIAL VACANT LAND
##                      464
##          32 HOSPITAL AND HEALTH FACILITIES
##                      38
##          33 EDUCATIONAL FACILITIES
##                      63
##          34 THEATRES
##                      9
## 35 INDOOR PUBLIC AND CULTURAL FACILITIES
##                      28
##          36 OUTDOOR RECREATIONAL FACILITIES
##                      14
##          37 RELIGIOUS FACILITIES
##                      102
##          38 ASYLUMS AND HOMES
##                      13
##          40 SELECTED GOVERNMENTAL FACILITIES
##                      7
##          41 TAX CLASS 4 - OTHER
##                      191
##                      NA's
##                      4204

```

```
bdset$borough <- as.factor(bdset$borough)
summary(bdset)
```

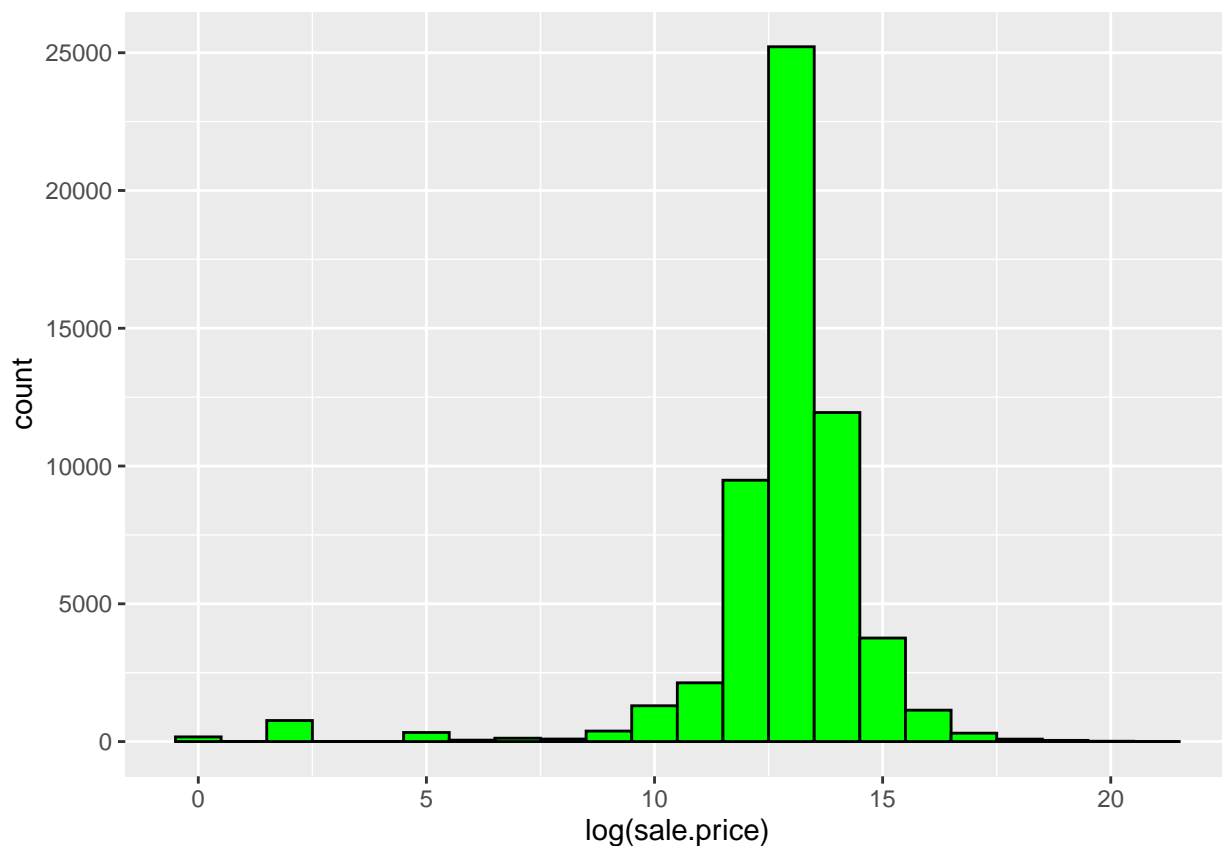
```
##           borough      neighborhood
## bronx      : 5268      Length:85975
## brooklyn   :23373      Class :character
## manhattan  :27395      Mode  :character
## queens     :23583
## statenisland: 6356
##
##
##           building.class.category tax.class.at.present
## 01 ONE FAMILY HOMES                :14846      Length:85975
## 10 COOPS - ELEVATOR APARTMENTS :13771      Class :character
## 02 TWO FAMILY HOMES                :13678      Mode  :character
## 13 CONDOS - ELEVATOR APARTMENTS:13313
## 03 THREE FAMILY HOMES                : 4135
## (Other)                             :22028
## NA's                                : 4204
##      block      lot      ease.ment      building.class.at.present
## Min.   :    1   Min.   :   1.0   Length:85975   Length:85975
## 1st Qu.: 1052   1st Qu.:  23.0   Class :character   Class :character
## Median : 2157   Median :   51.0   Mode  :character   Mode  :character
## Mean   : 3661   Mean   : 405.4
## 3rd Qu.: 5599   3rd Qu.:1009.0
## Max.   :16323   Max.   :9117.0
##
##      address      apart.ment.number      zip.code      residential.units
## Length:85975      Length:85975      Min.   :    0   Min.   :   0.00
## Class :character   Class :character   1st Qu.:10028   1st Qu.:   0.00
## Mode  :character   Mode  :character   Median :11201   Median :   1.00
##                                     Mean   :10758   Mean   :   1.96
##                                     3rd Qu.:11238   3rd Qu.:   2.00
##                                     Max.   :11694   Max.   :904.00
##                                     NA's    :3
## commercial.units   total.units      land.square.feet   gross.square.feet
## Min.   : 0.0000   Min.   : 0.000   Min.   :    0   Min.   : 0.00
## 1st Qu.: 0.0000   1st Qu.: 1.000   1st Qu.:    0   1st Qu.: 0.00
## Median : 0.0000   Median : 1.000   Median :  1512   Median : 0.00
## Mean   : 0.2349   Mean   : 2.252   Mean   :  3085   Mean   : 26.17
## 3rd Qu.: 0.0000   3rd Qu.: 2.000   3rd Qu.:  2625   3rd Qu.: 0.00
## Max.   :604.0000   Max.   :904.000   Max.   :7446955   Max.   :999.00
## NA's    :1        NA's    :4        NA's    :41789
## year.built   tax.class.at.time.of.sale   building.class.at.time.of.sale
## Min.   :    0   Min.   :1.000      Length:85975
## 1st Qu.:1910   1st Qu.:1.000      Class :character
## Median :1931   Median :2.000      Mode  :character
## Mean   :1681   Mean   :1.868
## 3rd Qu.:1964   3rd Qu.:2.000
## Max.   :2013   Max.   :4.000
##                                     NA's    :1
## sale.price      sale.date
## Min.   :    0   Min.   :2012-08-01
```

```
## 1st Qu.:      0    1st Qu.:2012-11-07
## Median : 260000    Median :2013-01-24
## Mean   : 885098    Mean   :2013-01-30
## 3rd Qu.: 620000    3rd Qu.:2013-05-02
## Max.   :1307965050    Max.   :2013-08-26
##
```

Outlier in sales price in log scale can be observed based on IQR and histogram. i.e Sale price < 10 and Sale price > 15

```
bdset1 <- bdset
ggplot(bdset1, mapping = aes(x = log(sale.price))) + geom_histogram(binwidth = 1,
  fill = "green", color = "black")
```

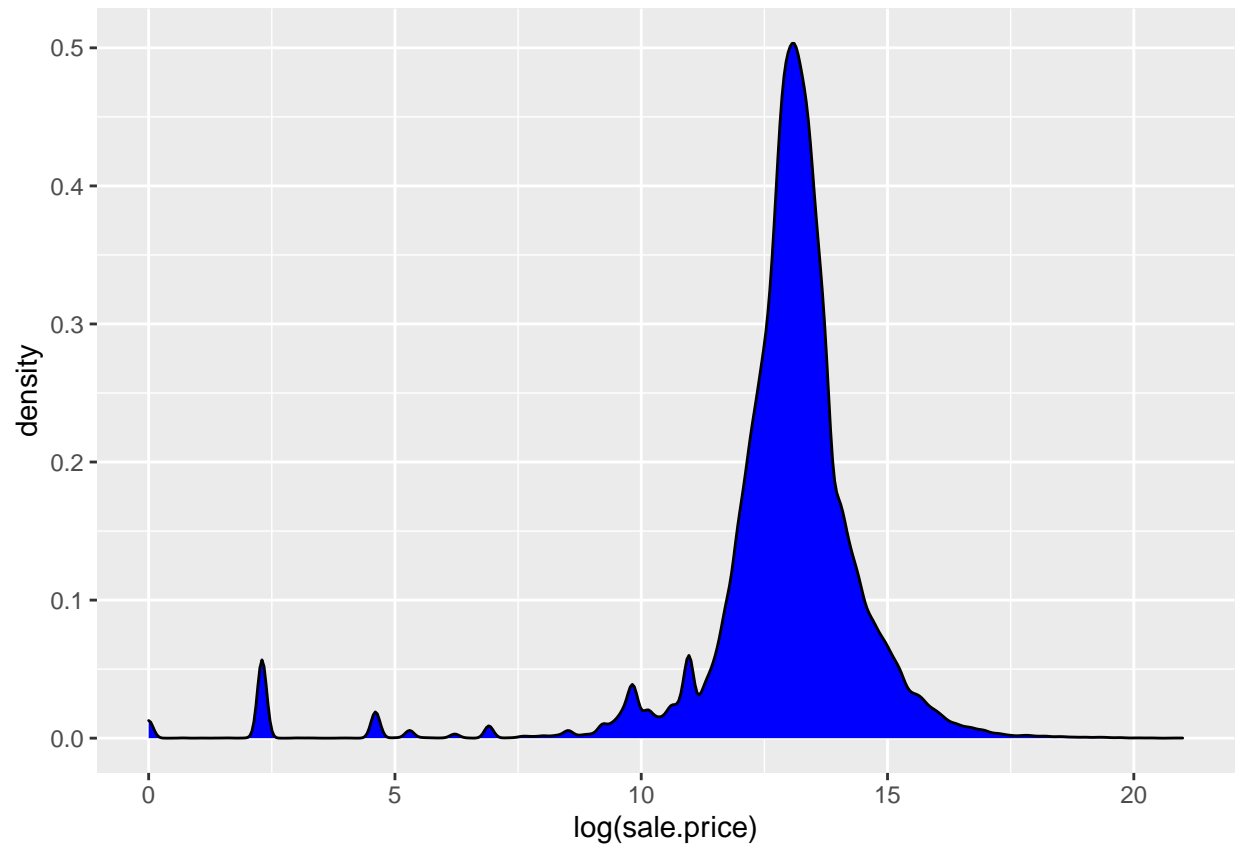
```
## Warning: Removed 28638 rows containing non-finite values (stat_bin).
```



```
ggplot(bdset1, mapping = aes(x = log(sale.price))) + geom_density(fill = "blue")
```

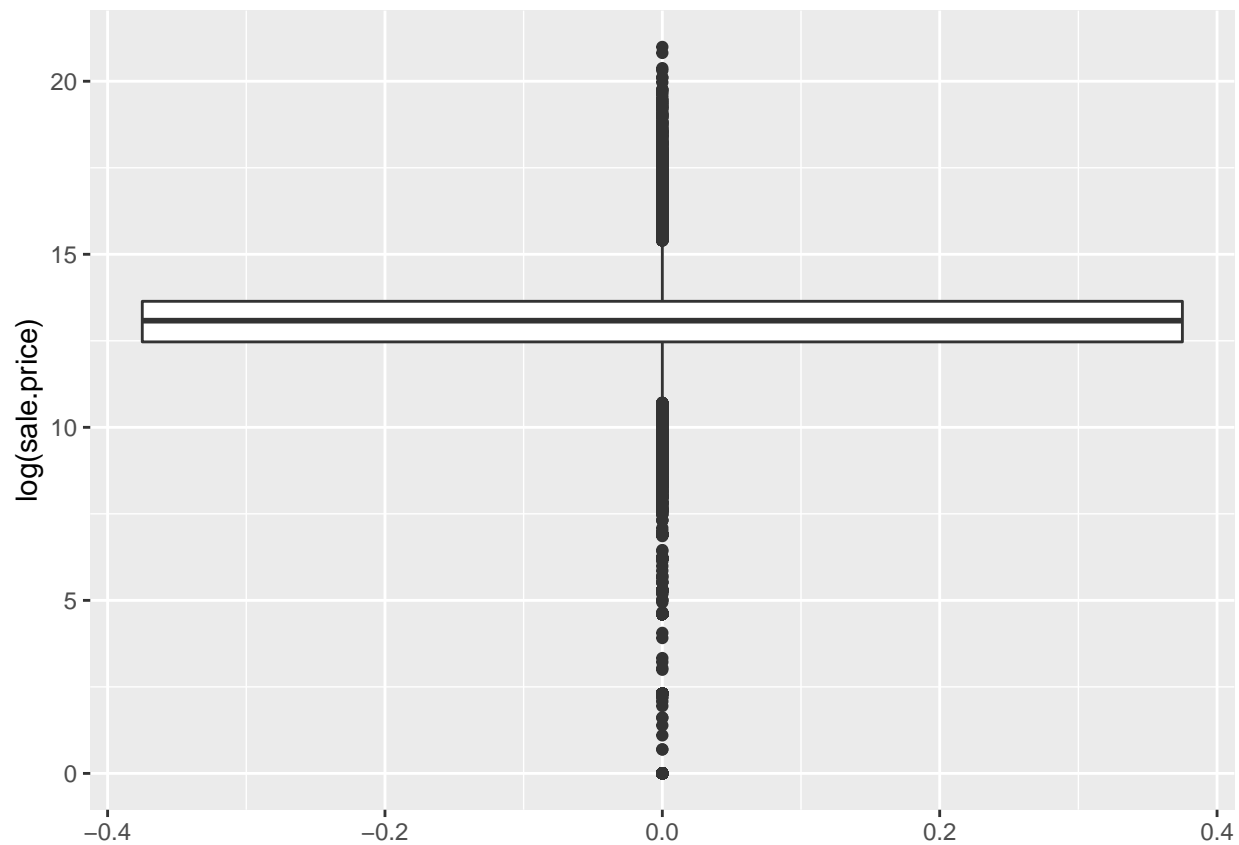
```
## Warning: Removed 28638 rows containing non-finite values (stat_density).
```





```
ggplot(bdset1) + geom_boxplot(mapping = aes(y = log(sale.price)))
```

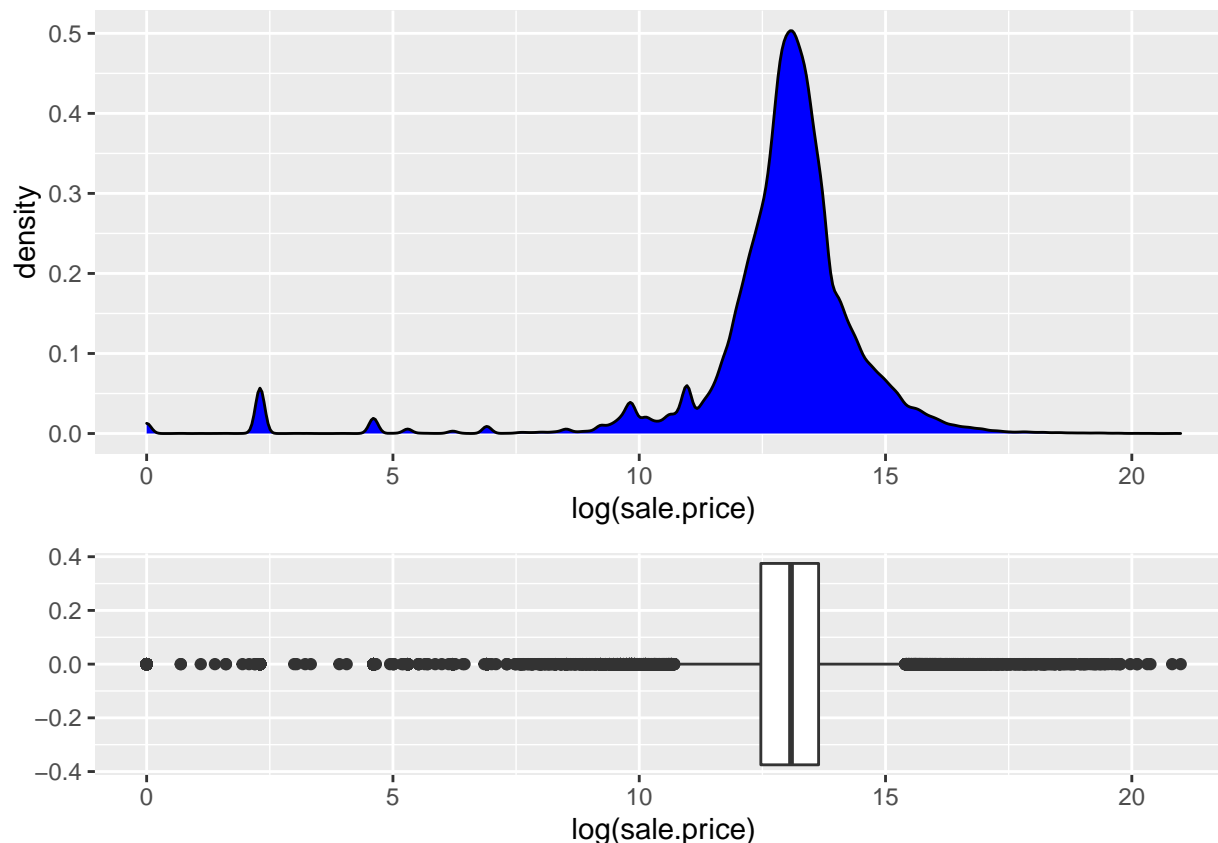
```
## Warning: Removed 28638 rows containing non-finite values (stat_boxplot).
```



```
egg::ggarrange(ggplot(bdset1, mapping = aes(x = log(sale.price))) + geom_density(fill = "blue"),
  ggplot(bdset1) + geom_boxplot(mapping = aes(x = log(sale.price))), heights = 2:1)
```

```
## Warning: Removed 28638 rows containing non-finite values (stat_density).
```

```
## Warning: Removed 28638 rows containing non-finite values (stat_boxplot).
```



Filtering categories for 1/2/3 Family Home, Coops and condos. Filtering outlier on sale.price and required building class categories

```
bdset1 <- bdset %>% mutate(sale.date.mw = floor_date(sale.date, "month"))
bdset1 <- bdset1 %>% mutate(buildingclasscat = case_when(building.class.category ==
  "01 ONE FAMILY HOMES" ~ "1 FH", building.class.category == "02 TWO FAMILY HOMES" ~
  "2 FH", building.class.category == "03 THREE FAMILY HOMES" ~ "3 FH", grepl("COOPS",
  building.class.category) == TRUE ~ "COOPS", grepl("CONDOS", building.class.category) ==
  TRUE ~ "CONDOS", TRUE ~ "NA"))
bdset1$buildingclasscat <- as.factor(bdset1$buildingclasscat)
bdset2 <- bdset1 %>% filter(log(sale.price) > 5, log(sale.price) <= 15, buildingclasscat %in%
  c("1 FH", "2 FH", "3 FH", "COOPS", "CONDOS"))
summary(bdset2)
```

```
##      borough      neighborhood
##  bronx      : 2715   Length:45243
##  brooklyn   :11845   Class :character
##  manhattan  :13339   Mode  :character
##  queens     :13916
##  statenisland: 3428
##
##
##      building.class.category tax.class.at.present
##  10 COOPS - ELEVATOR APARTMENTS :12077   Length:45243
##  01 ONE FAMILY HOMES           : 9322     Class :character
##  13 CONDOS - ELEVATOR APARTMENTS: 8500     Mode  :character
```

```

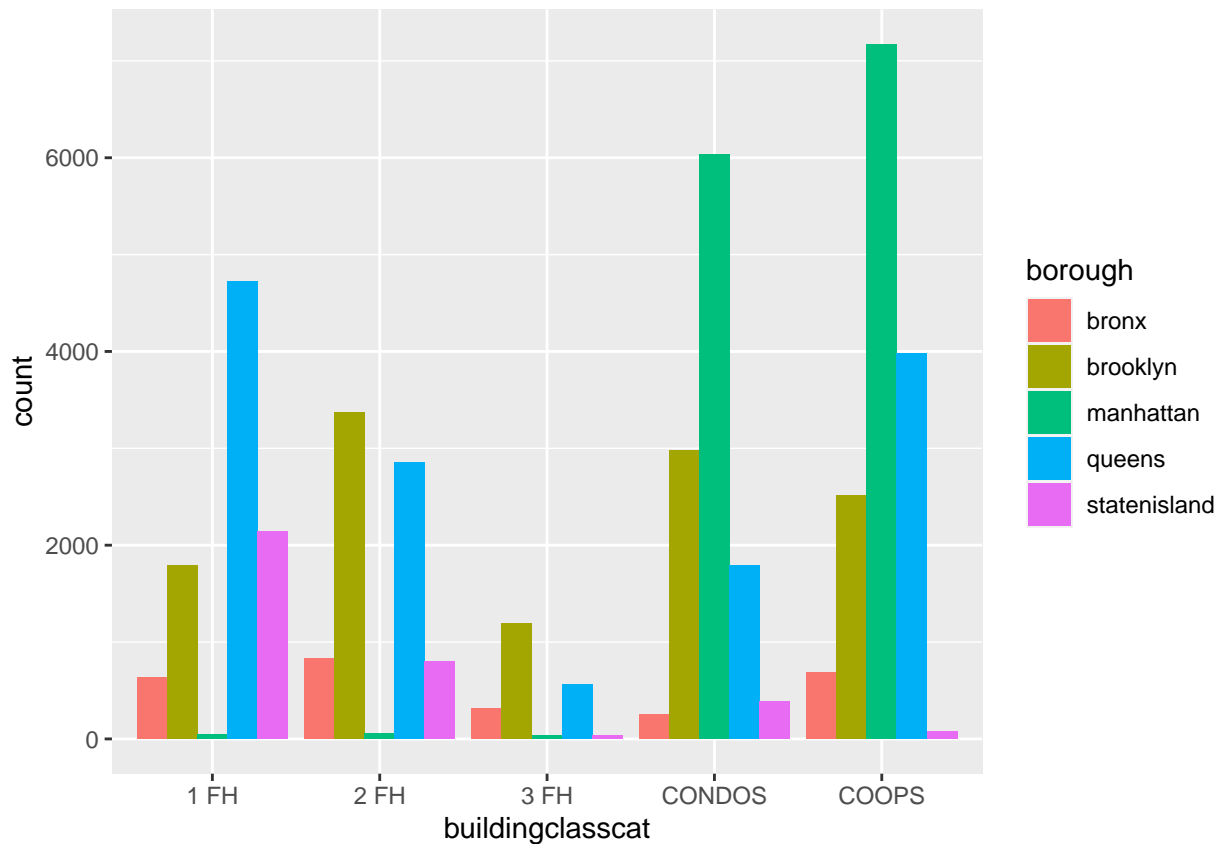
## 02 TWO FAMILY HOMES : 7916
## 09 COOPS - WALKUP APARTMENTS : 2346
## 03 THREE FAMILY HOMES : 2137
## (Other) : 2945
##      block      lot      ease.ment      building.class.at.present
## Min.   :    1   Min.   :    1.0   Length:45243   Length:45243
## 1st Qu.: 1179   1st Qu.:   20.0   Class :character   Class :character
## Median : 2298   Median :   49.0   Mode  :character   Mode  :character
## Mean   : 3853   Mean    : 379.7
## 3rd Qu.: 5913   3rd Qu.:1002.0
## Max.   :16320   Max.    :9034.0
##
##      address      apart.ment.number      zip.code      residential.units
## Length:45243      Length:45243      Min.   :    0   Min.   :    0.000
## Class :character   Class :character   1st Qu.:10038   1st Qu.:    0.000
## Mode  :character   Mode  :character   Median :11206   Median :    1.000
##                                     Mean   :10799   Mean    :    1.052
##                                     3rd Qu.:11356   3rd Qu.:    1.000
##                                     Max.   :11694   Max.    :437.000
##
##      commercial.units      total.units      land.square.feet      gross.square.feet
## Min.   :0.00000   Min.   :    0.000   Min.   :    0   Min.   :    0.00
## 1st Qu.:0.00000   1st Qu.:    0.000   1st Qu.:    0   1st Qu.:    0.00
## Median :0.00000   Median :    1.000   Median :    0   Median :    0.00
## Mean   :0.01837   Mean    :    1.079   Mean    : 1339   Mean    : 26.31
## 3rd Qu.:0.00000   3rd Qu.:    1.000   3rd Qu.: 2242   3rd Qu.:    0.00
## Max.   :9.00000   Max.    :437.000   Max.    :333700   Max.    :999.00
##                                     NA's    :18555
##      year.built      tax.class.at.time.of.sale      building.class.at.time.of.sale
## Min.   :    0   Min.   :1.000      Length:45243
## 1st Qu.:1920   1st Qu.:1.000      Class :character
## Median :1940   Median :2.000      Mode  :character
## Mean   :1797   Mean    :1.577
## 3rd Qu.:1963   3rd Qu.:2.000
## Max.   :2012   Max.    :4.000
##                                     NA's    :1
##      sale.price      sale.date      sale.date.mw      buildingclasscat
## Min.   :    200   Min.   :2012-08-01   Min.   :2012-08-01   1 FH : 9322
## 1st Qu.: 285000   1st Qu.:2012-11-07   1st Qu.:2012-11-01   2 FH : 7916
## Median : 469000   Median :2013-02-07   Median :2013-02-01   3 FH : 2137
## Mean   : 618969   Mean    :2013-02-06   Mean    :2013-01-21   CONDOS:11445
## 3rd Qu.: 745000   3rd Qu.:2013-05-15   3rd Qu.:2013-05-01   COOPS :14423
## Max.   :3262673   Max.    :2013-08-26   Max.    :2013-08-01   NA    :    0
##

```

## Data analysis

Plotting bar chart of building class category against borough, we can see number of Condos and Coops were sold more in Manhattan compared to other boroughs. Queens has highest number of single family homes sold while two family homes are were sold more in brooklyn and with queens beeing the next highest. Three family homes sales were not as much as the other categories

```
ggplot(bdset2) + geom_bar(mapping = aes(x = buildingclasscat, fill = borough), position = "dodge")
```



Line plot we can see that

1. Condo sale price spiked after July in bronx
2. Coops sale price remain consistent throughout the year in Brooklyn, Manhattan and Queens. Price fluctuations every quarter in Bronx and Staten Island
3. Single family home prices dipped in Manhattan and Bronx for Sep 2012. Overall the prices were consistent for rest of they months
4. Three family home prices were sold for less in October 2012, Feb 2012 and April 2012
5. Sale price fluctuates throught the year for Two family homes in Manhattan

```
ggplot(bdset2) + stat_summary(aes(x = sale.date.mw, y = log(sale.price), group = borough,
  color = borough), fun = mean, geom = "line") + scale_x_date(NULL, date_breaks = "3 month",
  date_labels = "%b%y") + facet_wrap(~buildingclasscat, ncol = 2)
```

