Keerthi CH Week2 Assignment

2022-07-15

Introduction

For this assignment i have worked on the provided zillow_price.csv dataset. Firstly, installed the necessary libraries and setup our working directory and loading the data and converting the file to a data table, so that it will be easier to look into the data.

```r
# load the data.table, ggolot2, and dplyr libraries and the zillow_price.csv file

library(data.table)
library(dplyr)
```

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

## The following objects are masked from 'package:data.table':
##
##     between, first, last

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

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

```r
library(ggplot2)

setwd("~/Downloads")
dt <- read.csv("zillow_price.csv")

# Convert the file to a data table
dt <- as.data.frame(dt)
```

Methods and Results

In this we are performing some histograms, boxplots, correlation, ggplot, Pearson correlation, linear model

Here we are seeing how many columns we have in our data set and also the summary and structure of the object.

```r
# how many observations and columns are there?
dim(dt)
```

```
## [1] 90275    60
```

```r
ncol(dt)
```

```
## [1] 60
```

```r
# use str and summary to see how many missing values we have,
# and what the data looks like
str(dt)
```

```
## 'data.frame':    90275 obs. of  60 variables:
##  $ parcelid                    : int  10711738 10711755 10711805 10711816 10711858 10711910 10712086
##  $ airconditioningtypeid       : int  1 1 1 1 1 NA 1 1 1 1 ...
##  $ architecturalstyletypeid    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ basementsqft                : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ bathroomcnt                 : num  3 3 2 2 2 2 2 3 3 3 ...
##  $ bedroomcnt                  : int  4 3 3 4 4 3 4 3 4 3 ...
##  $ buildingclasstypeid         : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ buildingqualitytypeid       : int  4 4 4 4 4 4 4 4 4 4 ...
##  $ calculatedbathnbr           : num  3 3 2 2 2 2 2 3 3 3 ...
##  $ decktypeid                  : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ finishedfloor1squarefeet    : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ calculatedfinishedsquarefeet: int  2538 1589 2411 2232 1882 1477 1850 3193 2421 1678 ...
##  $ finishedsquarefeet12        : int  2538 1589 2411 2232 1882 1477 1850 3193 2421 1678 ...
##  $ finishedsquarefeet13        : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ finishedsquarefeet15        : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ finishedsquarefeet50        : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ finishedsquarefeet6         : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ fips                        : int  6037 6037 6037 6037 6037 6037 6037 6037 6037 6037 ...
##  $ fireplacecnt                : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ fullbathcnt                 : int  3 3 2 2 2 2 2 3 3 3 ...
##  $ garagecarcnt                : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ garagetotalsqft             : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ hashottuborspa              : chr  "" "" "" "" ...
##  $ heatingorsystemtypeid       : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ latitude                    : int  34220381 34222040 34220427 34222390 34222544 34221864 34226039
##  $ longitude                   : int  -118620802 -118622240 -118618549 -118618631 -118617961 -118615
##  $ lotsizesquarefeet           : num  11012 11010 11723 9002 9002 ...
##  $ poolcnt                     : int  1 1 1 NA 1 1 1 1 1 NA ...
##  $ poolsizesum                 : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ pooltypeid10                : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ pooltypeid2                 : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ pooltypeid7                 : int  1 1 1 NA 1 1 1 1 1 NA ...
##  $ propertycountylandusecode   : chr  "0101" "0101" "0101" "0100" ...
##  $ propertylandusetypeid       : int  261 261 261 261 261 261 261 261 261 261 ...
##  $ propertyzoningdesc          : chr  "LARE11" "LARE11" "LARE9" "LARE9" ...
##  $ rawcensustractandblock      : num  60371132 60371132 60371132 60371132 60371132 ...
##  $ regionidcity                : int  12447 12447 12447 12447 12447 12447 12447 12447 12447 12447 ..
##  $ regionidcounty              : int  3101 3101 3101 3101 3101 3101 3101 3101 3101 3101 ...
##  $ regionidneighborhood        : int  268588 268588 268588 268588 268588 268588 268588 268588 268588
##  $ regionidzip                 : int  96339 96339 96339 96339 96339 96339 96339 96339 96339 96339 ..
##  $ roomcnt                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ storytypeid                 : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ threequarterbathnbr         : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ typeconstructiontypeid      : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ unitcnt                     : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ yardbuildingsqft17          : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ yardbuildingsqft26          : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ yearbuilt                   : int  1978 1959 1973 1973 1973 1960 1974 1964 1962 1961 ...
##  $ numberofstories             : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ fireplaceflag               : chr  "" "" "" "" ...
##  $ structuretaxvaluedollarcnt  : num  245180 254691 235114 262309 232037 ...
##  $ taxvaluedollarcnt           : num  567112 459844 384787 437176 382055 ...
##  $ assessmentyear              : int  2015 2015 2015 2015 2015 2015 2015 2015 2015 2015 ...
##  $ landtaxvaluedollarcnt       : num  321932 205153 149673 174867 150018 ...
##  $ taxdelinquencyflag          : chr  "" "" "" "" ...
##  $ taxdelinquencyyear          : int  NA NA NA NA NA NA NA NA NA NA ...
##  $ censustractandblock         : num  6.04e+13 6.04e+13 6.04e+13 6.04e+13 6.04e+13 ...
##  $ price                       : num  622343 594922 420397 479316 420539 ...
##  $ logerror                    : num  0.0276 -0.0182 -0.1009 -0.0121 -0.0481 ...
##  $ transactiondate             : chr  "2016-08-02" "2016-08-02" "2016-05-03" "2016-04-05" ...
```

```
summary(dt)
```

```
##     parcelid         airconditioningtypeid architecturalstyletypeid
##  Min.   : 10711738   Min.   : 1.00         Min.   : 2.00
##  1st Qu.: 11559500   1st Qu.: 1.00         1st Qu.: 7.00
##  Median : 12547337   Median : 1.00         Median : 7.00
##  Mean   : 12984656   Mean   : 1.82         Mean   : 7.23
##  3rd Qu.: 14227552   3rd Qu.: 1.00         3rd Qu.: 7.00
##  Max.   :162960842   Max.   :13.00         Max.   :21.00
##                      NA's   :61494         NA's   :90014
##   basementsqft     bathroomcnt       bedroomcnt     buildingclasstypeid
##  Min.   : 100.0   Min.   : 0.000   Min.   : 0.000   Min.   :4
##  1st Qu.: 407.5   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.:4
##  Median : 616.0   Median : 2.000   Median : 3.000   Median :4
##  Mean   : 713.6   Mean   : 2.279   Mean   : 3.032   Mean   :4
##  3rd Qu.: 872.0   3rd Qu.: 3.000   3rd Qu.: 4.000   3rd Qu.:4
##  Max.   :1555.0   Max.   :20.000   Max.   :16.000   Max.   :4
##  NA's   :90232                                      NA's   :90259
##  buildingqualitytypeid calculatedbathnbr   decktypeid
##  Min.   : 1.00         Min.   : 1.000    Min.   :66
##  1st Qu.: 4.00         1st Qu.: 2.000    1st Qu.:66
##  Median : 7.00         Median : 2.000    Median :66
##  Mean   : 5.57         Mean   : 2.309    Mean   :66
##  3rd Qu.: 7.00         3rd Qu.: 3.000    3rd Qu.:66
##  Max.   :12.00         Max.   :20.000    Max.   :66
##  NA's   :32911         NA's   :1182      NA's   :89617
##  finishedfloor1squarefeet calculatedfinishedsquarefeet finishedsquarefeet12
##  Min.   :  44             Min.   :    2                Min.   :    2
##  1st Qu.: 938             1st Qu.: 1184                1st Qu.: 1172
##  Median :1244             Median : 1540                Median : 1518
##  Mean   :1348             Mean   : 1773                Mean   : 1745
##  3rd Qu.:1614             3rd Qu.: 2095                3rd Qu.: 2056
##  Max.   :7625             Max.   :22741                Max.   :20013
##  NA's   :83419            NA's   :661                  NA's   :4679
##  finishedsquarefeet13 finishedsquarefeet15 finishedsquarefeet50
##  Min.   :1056         Min.   :  560        Min.   :  44
##  1st Qu.:1392         1st Qu.: 1648        1st Qu.: 938
##  Median :1440         Median : 2104        Median :1248
##  Mean   :1405         Mean   : 2380        Mean   :1356
##  3rd Qu.:1440         3rd Qu.: 2862        3rd Qu.:1619
##  Max.   :1584         Max.   :22741        Max.   :8352
##  NA's   :90242        NA's   :86711        NA's   :83419
##  finishedsquarefeet6      fips       fireplacecnt    fullbathcnt
##  Min.   : 257        Min.   :6037   Min.   :1.00    Min.   : 1.000
##  1st Qu.:1112        1st Qu.:6037   1st Qu.:1.00    1st Qu.: 2.000
##  Median :2028        Median :6037   Median :1.00    Median : 2.000
##  Mean   :2303        Mean   :6049   Mean   :1.19    Mean   : 2.241
##  3rd Qu.:3431        3rd Qu.:6059   3rd Qu.:1.00    3rd Qu.: 3.000
##  Max.   :7224        Max.   :6111   Max.   :5.00    Max.   :20.000
##  NA's   :89854                       NA's   :80668   NA's   :1182
##   garagecarcnt   garagetotalsqft  hashottuborspa     heatingorsystemtypeid
##  Min.   : 0.00   Min.   :   0.0   Length:90275       Min.   : 1.00
##  1st Qu.: 2.00   1st Qu.:   0.0   Class :character   1st Qu.: 2.00
##  Median : 2.00   Median : 433.0   Mode  :character   Median : 2.00
##  Mean   : 1.81   Mean   : 345.5                      Mean   : 3.93
##  3rd Qu.: 2.00   3rd Qu.: 484.0                      3rd Qu.: 7.00
##  Max.   :24.00   Max.   :7339.0                      Max.   :24.00
##  NA's   :60338   NA's   :60338                       NA's   :34195
##     latitude          longitude          lotsizesquarefeet    poolcnt
##  Min.   :33339295   Min.   :-119447865   Min.   :    167   Min.   :1
##  1st Qu.:33811538   1st Qu.:-118411692   1st Qu.:   5703   1st Qu.:1
##  Median :34021500   Median :-118173431   Median :   7200   Median :1
##  Mean   :34005411   Mean   :-118198868   Mean   :  29110   Mean   :1
##  3rd Qu.:34172742   3rd Qu.:-117921588   3rd Qu.:  11686   3rd Qu.:1
##  Max.   :34816009   Max.   :-117554924   Max.   :6971010   Max.   :1
##                                          NA's   :10150     NA's   :72374
##   poolsizesum      pooltypeid10    pooltypeid2     pooltypeid7
##  Min.   :  28.0   Min.   :1       Min.   :1       Min.   :1
##  1st Qu.: 420.0   1st Qu.:1       1st Qu.:1       1st Qu.:1
##  Median : 500.0   Median :1       Median :1       Median :1
##  Mean   : 519.8   Mean   :1       Mean   :1       Mean   :1
##  3rd Qu.: 600.0   3rd Qu.:1       3rd Qu.:1       3rd Qu.:1
##  Max.   :1750.0   Max.   :1       Max.   :1       Max.   :1
##  NA's   :89306    NA's   :89114   NA's   :89071   NA's   :73578
##  propertycountylandusecode propertylandusetypeid propertyzoningdesc
##  Length:90275              Min.   : 31.0         Length:90275
##  Class :character          1st Qu.:261.0         Class :character
##  Mode  :character          Median :261.0         Mode  :character
##                            Mean   :261.8
##                            3rd Qu.:266.0
##                            Max.   :275.0
##
##  rawcensustractandblock  regionidcity    regionidcounty regionidneighborhood
##  Min.   :60371011       Min.   :  3491   Min.   :1286   Min.   :  6952
##  1st Qu.:60373203       1st Qu.: 12447   1st Qu.:1286   1st Qu.: 46736
##  Median :60376200       Median : 25218   Median :3101   Median :118887
##  Mean   :60491795       Mean   : 33761   Mean   :2525   Mean   :190646
##  3rd Qu.:60590423       3rd Qu.: 45457   3rd Qu.:3101   3rd Qu.:274800
##  Max.   :61110091       Max.   :396556   Max.   :3101   Max.   :764167
##                         NA's   :1803                    NA's   :54263
##   regionidzip        roomcnt        storytypeid    threequarterbathnbr
##  Min.   : 95982   Min.   : 0.000   Min.   :7       Min.   :1.00
##  1st Qu.: 96193   1st Qu.: 0.000   1st Qu.:7       1st Qu.:1.00
##  Median : 96393   Median : 0.000   Median :7       Median :1.00
##  Mean   : 96586   Mean   : 1.479   Mean   :7       Mean   :1.01
##  3rd Qu.: 96987   3rd Qu.: 0.000   3rd Qu.:7       3rd Qu.:1.00
##  Max.   :399675   Max.   :18.000   Max.   :7       Max.   :4.00
##  NA’s   :35                        NA’s   :90232   NA’s   :78266
##  typeconstructiontypeid    unitcnt       yardbuildingsqft17 yardbuildingsqft26
##  Min.   : 4.00          Min.   :  1.00   Min.   :  25.0     Min.   :  18.0
##  1st Qu.: 6.00          1st Qu.:  1.00   1st Qu.: 180.0     1st Qu.: 100.0
##  Median : 6.00          Median :  1.00   Median : 259.5     Median : 159.0
##  Mean   : 6.01          Mean   :  1.11   Mean   : 310.1     Mean   : 311.7
##  3rd Qu.: 6.00          3rd Qu.:  1.00   3rd Qu.: 384.0     3rd Qu.: 361.0
##  Max.   :13.00          Max.   :143.00   Max.   :2678.0     Max.   :1366.0
##  NA’s   :89976          NA’s   :31922    NA’s   :87629      NA’s   :90180
##    yearbuilt    numberofstories fireplaceflag      structuretaxvaluedollarcnt
##  Min.   :1885   Min.   :1.00    Length:90275       Min.   :    100
##  1st Qu.:1953   1st Qu.:1.00    Class :character   1st Qu.:  81245
##  Median :1970   Median :1.00    Mode  :character   Median : 132000
##  Mean   :1969   Mean   :1.44                       Mean   : 180093
##  3rd Qu.:1987   3rd Qu.:2.00                       3rd Qu.: 210534
##  Max.   :2015   Max.   :4.00                       Max.   :9948100
##  NA’s   :756    NA’s   :69705                      NA’s   :380
##  taxvaluedollarcnt  assessmentyear landtaxvaluedollarcnt taxdelinquencyflag
##  Min.   :      22   Min.   :2015   Min.   :      22      Length:90275
##  1st Qu.:  199023   1st Qu.:2015   1st Qu.:   82228      Class :character
##  Median :  342872   Median :2015   Median :  192970      Mode  :character
##  Mean   :  457673   Mean   :2015   Mean   :  278335
##  3rd Qu.:  540589   3rd Qu.:2015   3rd Qu.:  345420
##  Max.   :27750000   Max.   :2015   Max.   :24500000
##  NA’s   :1                         NA’s   :1
##  taxdelinquencyyear censustractandblock     price            logerror
##  Min.   : 6.0       Min.   :6.037e+13   Min.   :    4231   Min.   :-4.60500
##  1st Qu.:13.0       1st Qu.:6.037e+13   1st Qu.:  247658   1st Qu.:-0.02530
##  Median :14.0       Median :6.038e+13   Median :  391616   Median : 0.00600
##  Mean   :13.4       Mean   :6.049e+13   Mean   :  515860   Mean   : 0.01146
##  3rd Qu.:15.0       3rd Qu.:6.059e+13   3rd Qu.:  594922   3rd Qu.: 0.03920
##  Max.   :99.0       Max.   :6.111e+13   Max.   :27753111   Max.   : 4.73700
##  NA’s   :88492      NA’s   :605         NA’s   :6
##  transactiondate
##  Length:90275
##  Class :character
##  Mode  :character
##
##
##
##
```

The results shows that we have 90,275 observations and 60 columns in the data table. And the summary gives the information about the number of Null values in each column.

In this next step we are taking the columns which are numeric and don’t have lot of missing values.

```r
# columns that are numeric and don't have lots of missing values
# you can add others if you like
numeric_cols <- c('bathroomcnt',
                  'bedroomcnt',
                  'calculatedfinishedsquarefeet',
                  'roomcnt',
                  'yearbuilt',
                  'taxvaluedollarcnt',
                  'landtaxvaluedollarcnt',
                  'price')

dt_rows = nrow(dt)

# Simplify your dataset by only selecting the columns of your choosing dt[, numeric_cols, with = FALSE]

dt <- dt[, numeric_cols]
dt <- na.omit(dt)
#dt <- dt[complete.cases(dt),]
dt_rows - nrow(dt)
```

```
## [1] 776
```

So, here we see that we have dropped 776 rows which has Null values.

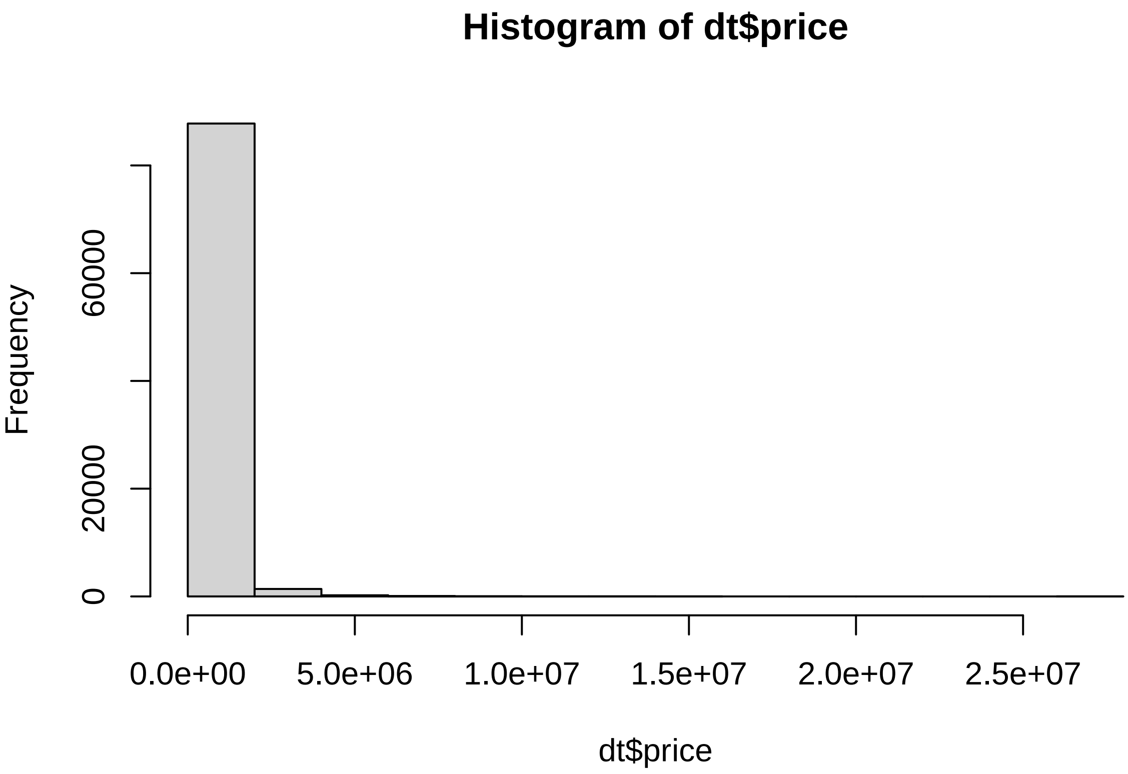
Here we are checking the correlation for home price and the taxvaluedollarcnt and i see that the correlation was found to be 0.95

Created a boxplot and histogram of the price data.

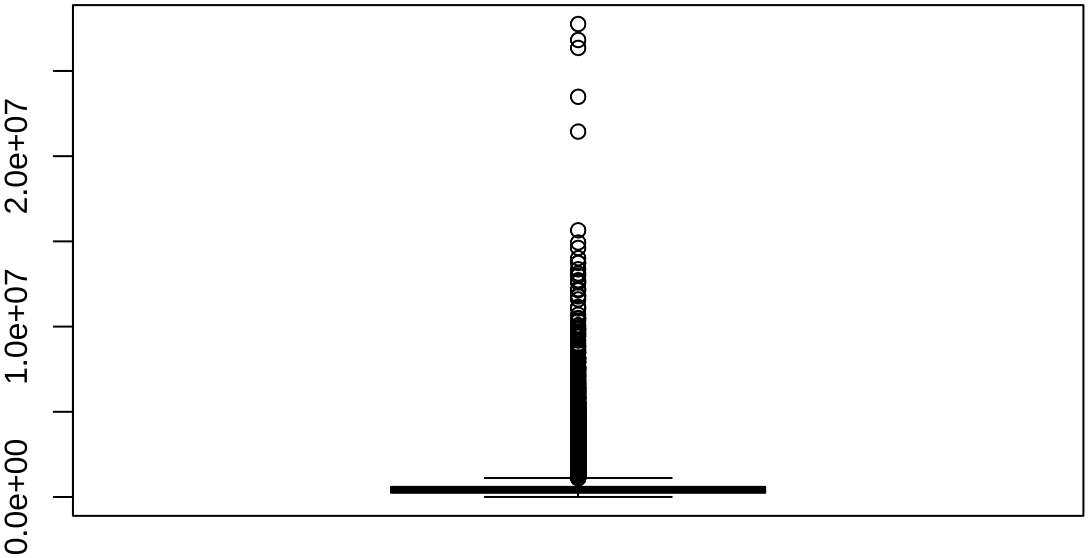
```r
# We want to try to correlate home price with another variable.
# Create a boxplot of the price data

cor(dt$price, dt$taxvaluedollarcnt)
```

```
## [1] 0.9518683
```

```r
hist(dt$price)
```

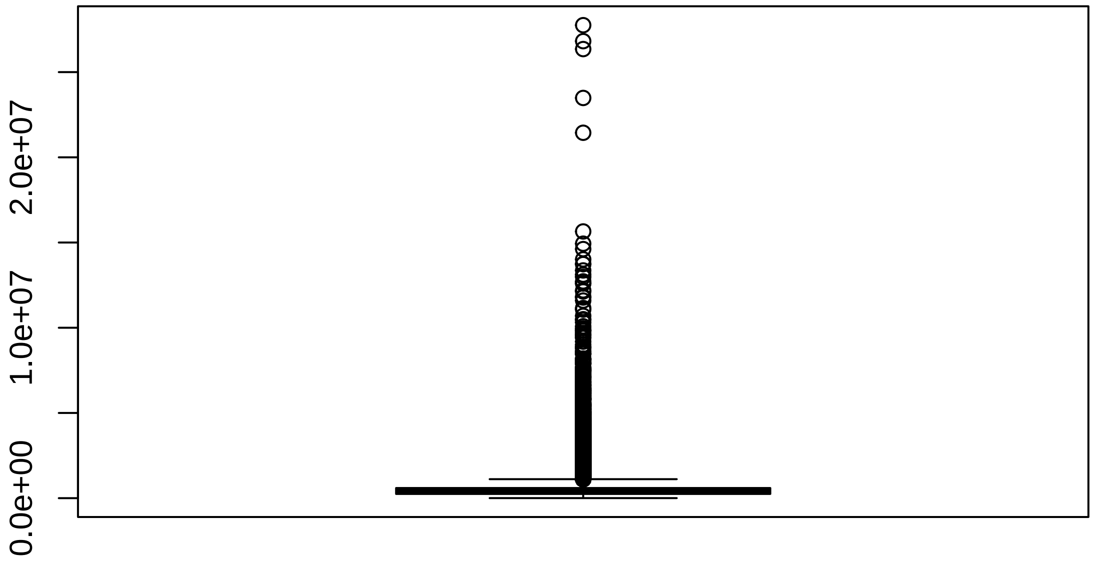


```
boxplot(dt$price)
```

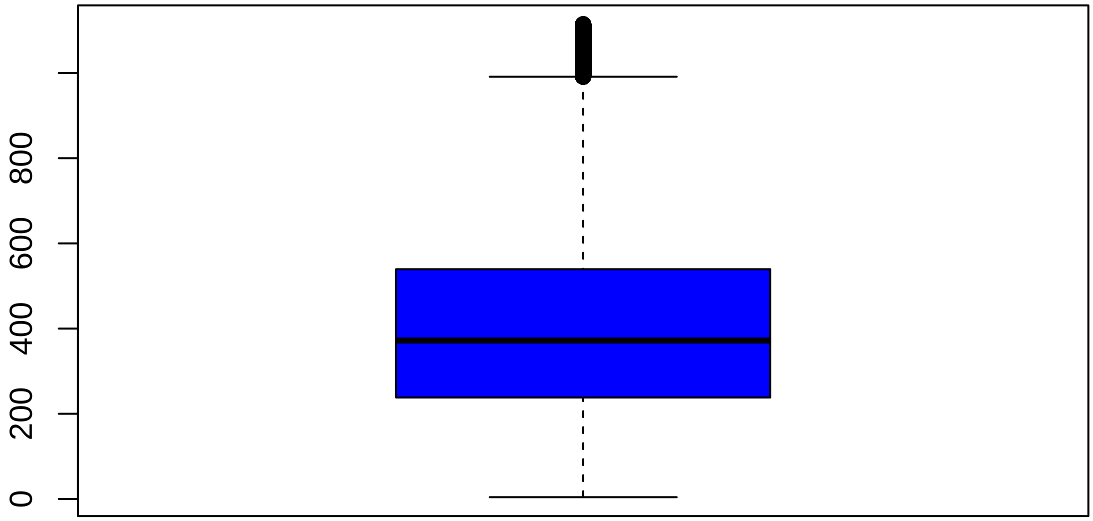


Here we are removing the outliers.

```
# Remove the outliers. dt[!which(dt$price %nin% boxplot(dt$price)$out)]

`%nin%` <- Negate(`%in%`)

dt1 <- dt[which(dt$price %nin% boxplot(dt$price)$out),]
```



```
`%nin%` <- Negate(`%in%`)
dt1 <- dt1 %>%
  mutate(price = price / 1000)

dt1 <- dt1 %>%
  mutate(taxvaluedollarcnt = taxvaluedollarcnt / 1000)


boxplot(dt1$price,
        col = "blue")
```



The boxplot shows us that there are houses ranging from very cheap to around $1M.

```
# How many outliers did we drop?
nrow(dt) - nrow(dt1)
```

```
## [1] 6020
```

Here it looks like we have dropped 6020 outliers.

As there are many observations lets take 200 samples and do some plotting.

```
# In our case, we have too many observations.
# Use sample() to only sample a few hundred points to plot.

dt_sample <- dt1[sample(nrow(dt1), size = 200),]
```

Here i have selected the variables price & taxvaluedollarcnt and performed the correlation..

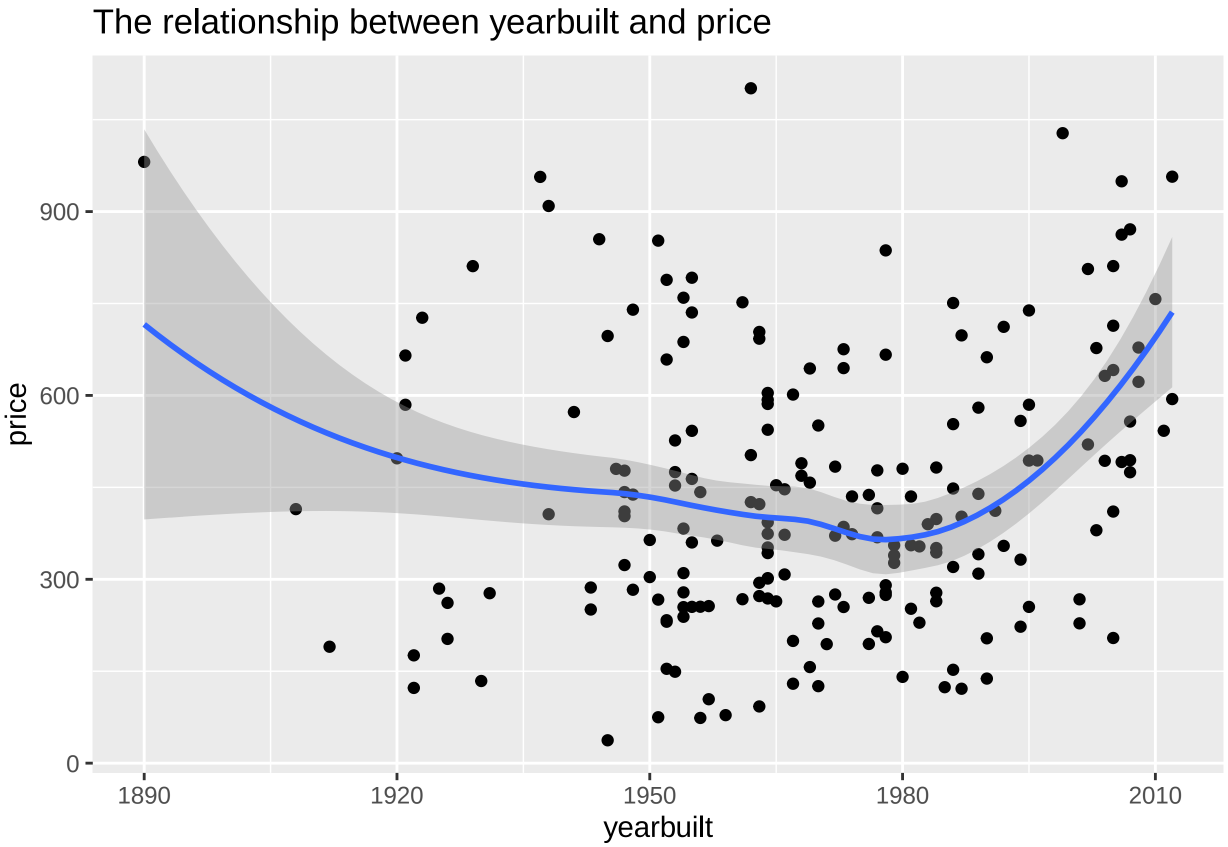
```
# plot a few of the more interesting pairs together
cor(dt1$price, dt1$taxvaluedollarcnt)
```

```
## [1] 0.9108692
```

From the below results we can see that the correlation for price and taxvaluedollarcnt was found to be 0.91.

```
# bonus: try to make some nice-looking scatter plots with ggplot2

ggplot(dt_sample, aes(yearbuilt, price)) +
geom_point() +
labs( title = "The relationship between yearbuilt and price")+ geom_smooth()
```

```
## 'geom_smooth()' using method = 'loess' and formula 'y ~ x'
```



```
# create a new data.table by dropping any missing values
# use dim() to see how many cases we dropped
colSums(dt1==0)
```

```
##                  bathroomcnt                   bedroomcnt
##                          449                          705
## calculatedfinishedsquarefeet                      roomcnt
##                            0                        63760
##                    yearbuilt            taxvaluedollarcnt
##                            0                            0
##        landtaxvaluedollarcnt                        price
##                            0                            0
```

```
dim(dt1)
```

```
## [1] 83479     8
```

From the above results we can see that the roomcnt has more number of zeros compared to others.

Here we are performing the pearson correlation for price and taxvaluedollarcnt and we see that the correlation was found to be 0.90 which is a positive correlation.

```
# get the pearson correlation between price and another variable using cor()
#...there are other types of correlations
# try ?cor to see options, and try another correlation

cor(dt1$price, dt1$taxvaluedollarcnt, method = "spearman")
```

```
## [1] 0.9052994
```

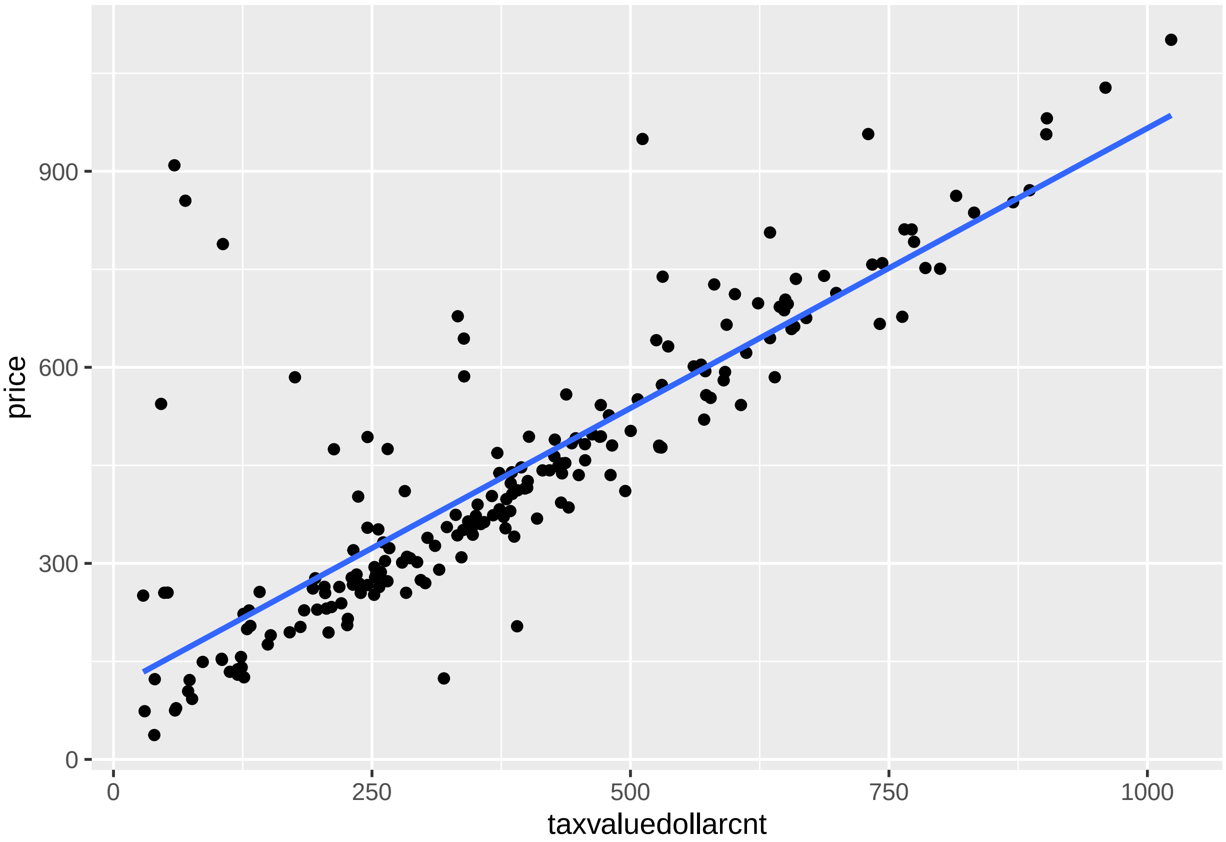
Created a linear model (lm) for a correlated variable by using the price and the taxvaluedollarcnt. Plotted the variables.

```
# use the lm() command to fit a linear model of price to the
# one variable you think is most correlated or predictive of price
# lm stands for 'linear model'

fit <- lm(dt_sample$taxvaluedollarcnt ~ dt_sample$price)

# view the model summary
summary(fit)
```

```
##
## Call:
## lm(formula = dt_sample$taxvaluedollarcnt ~ dt_sample$price)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -712.46  -24.36   19.47   55.41  199.08
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      23.79131   18.56174   1.282    0.201
## dt_sample$price   0.82231    0.03784  21.729   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 117.7 on 198 degrees of freedom
## Multiple R-squared:  0.7045, Adjusted R-squared:  0.703
## F-statistic: 472.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
# plot a scatter plot of the price and the variable you chose

ggplot(dt_sample, aes(taxvaluedollarcnt, price)) +
  geom_point() +
  geom_smooth(formula = y ~ x, method = 'lm', se = FALSE, data = dt_sample)
```

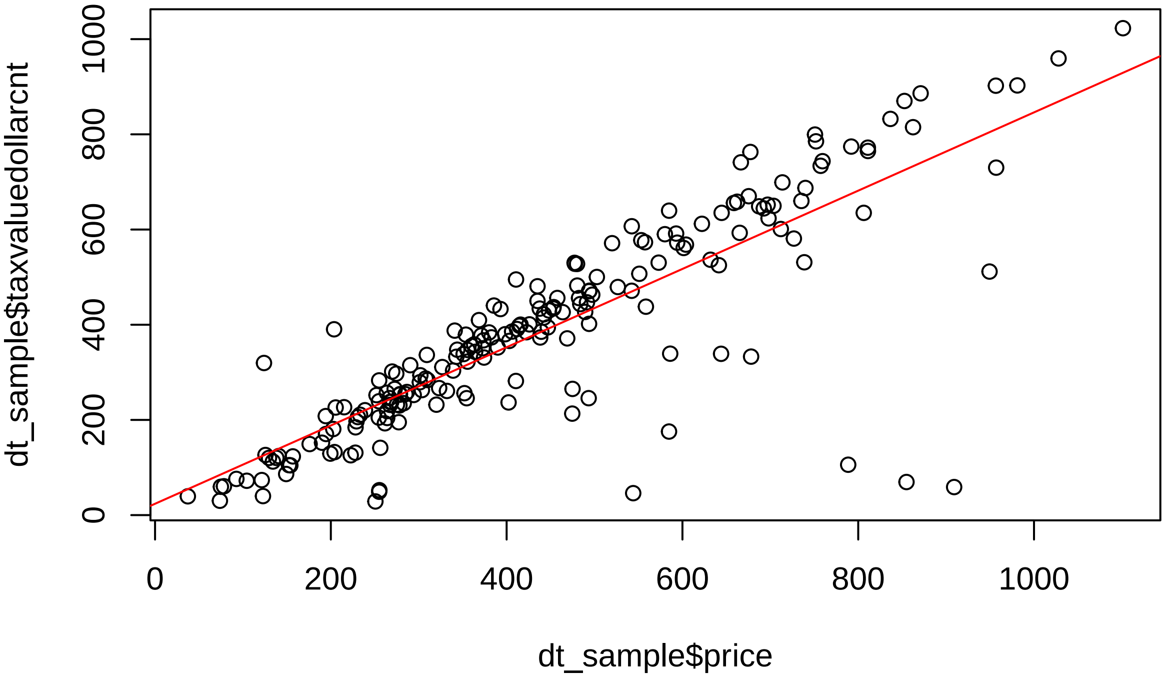


```
  labs( title = "Tax value to Price")
```

```
## $title
## [1] "Tax value to Price"
##
## attr(,"class")
## [1] "labels"
```

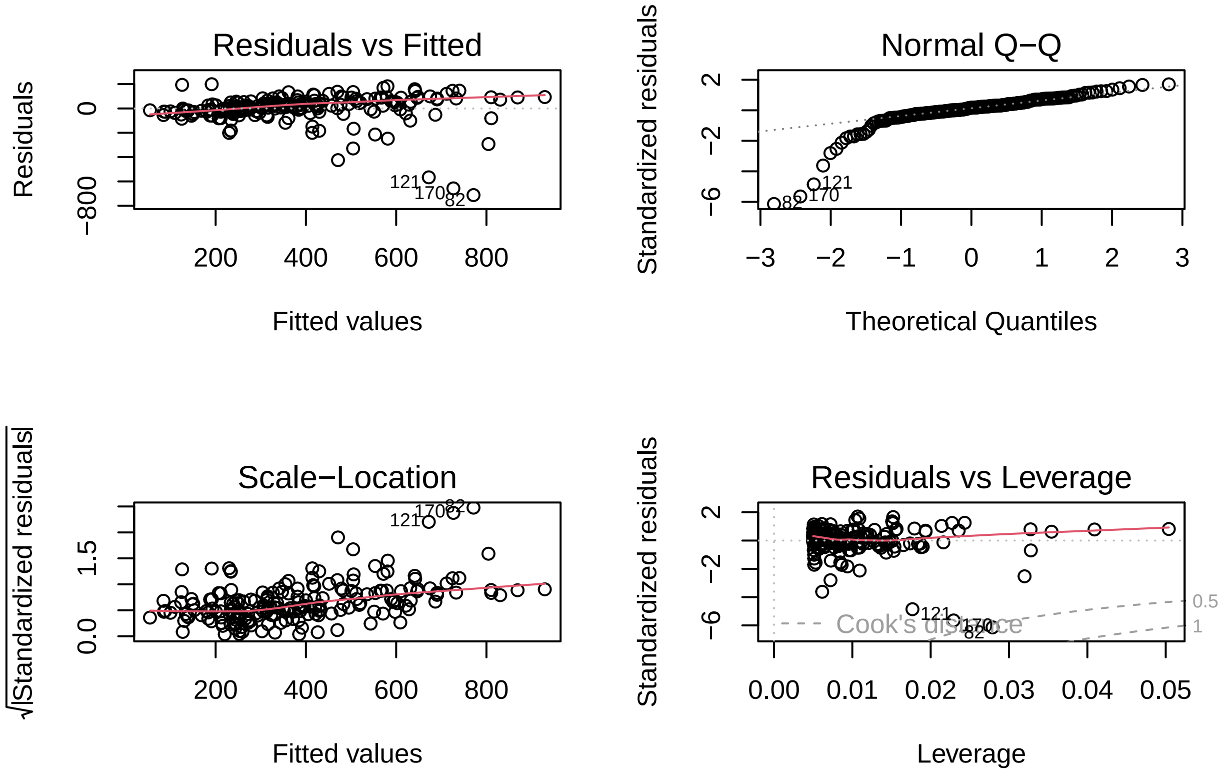
Added a regression line using abline and below is the plot.

```
# add the regression line to the current plot using abline()
plot(dt_sample$price, dt_sample$taxvaluedollarcnt)
abline(fit, col="red")
```



We can see that they are identical with the exception of the graphics, both indicating a good fit for the model. The ggplot2 model is definitely prettier but with some effort plot can be just as nice.

```
# plot the fit diagnostics here
par(mfrow=c(2,2)) # Change the panel layout to 2 x 2
plot(fit)
```



The Residual vs fitted plot can show if there is a non-linear relationship between our variables. But here we can see a flat, straight line indicating a linear relationship.

Normal Q-Q plot looks for normal distribution. we can see it is mostly normally distributed, but i think this indicates a slight skewing to the right where the lower priced houses are the exception.

The scale location plot shows the spread of residuals. A flat line would mean a uniform spread. However our data shows a higher spread as the value increases.

The Residuals vs Leverage plot shows extreme value cases can have on the regression line.

Conclusion:

I have downloaded the dataset and converted it to the data table so that it will be easier to look at the data, and i see that there are 90,275 observations and 60 columns in the data table and the summary gave the information about the number of Null values in each column. Next i have dropped 776 rows which has Null values.

Next i have checked the correlation for price and the taxvaluedollarcnt and i see that the correlation was found to be 0.95

Performed box plot to see the outliers in a price column and removed them, and after removing i have plotted the boxplot again and see that the there are houses ranging from very cheap to around $1M approximately.

Observation : I have performed the correlation between price and taxvaluedollarcnt using pearson model and see that my results are 0.90 which is really good..

Next i have selected price:Taxvaluedollarcnt and performed a basic linear model and performed a scatter plot and added a regression line using abline and we see that they are identical with the exception of graphics indicating a good fit for the model and see that we have a strong positive correlation, i also found that outliers do have a negative effect on our model.