project X

2023-02-28

####summary and visualization

summary(data)

## date price bedrooms bathrooms   
## Length:4600 Min. : 0 Min. :0.000 Min. :0.000   
## Class :character 1st Qu.: 322875 1st Qu.:3.000 1st Qu.:1.750   
## Mode :character Median : 460943 Median :3.000 Median :2.250   
## Mean : 551963 Mean :3.401 Mean :2.161   
## 3rd Qu.: 654962 3rd Qu.:4.000 3rd Qu.:2.500   
## Max. :26590000 Max. :9.000 Max. :8.000   
## sqft\_living sqft\_lot floors waterfront   
## Min. : 370 Min. : 638 Min. :1.000 Min. :0.000000   
## 1st Qu.: 1460 1st Qu.: 5001 1st Qu.:1.000 1st Qu.:0.000000   
## Median : 1980 Median : 7683 Median :1.500 Median :0.000000   
## Mean : 2139 Mean : 14852 Mean :1.512 Mean :0.007174   
## 3rd Qu.: 2620 3rd Qu.: 11001 3rd Qu.:2.000 3rd Qu.:0.000000   
## Max. :13540 Max. :1074218 Max. :3.500 Max. :1.000000   
## view condition sqft\_above sqft\_basement   
## Min. :0.0000 Min. :1.000 Min. : 370 Min. : 0.0   
## 1st Qu.:0.0000 1st Qu.:3.000 1st Qu.:1190 1st Qu.: 0.0   
## Median :0.0000 Median :3.000 Median :1590 Median : 0.0   
## Mean :0.2407 Mean :3.452 Mean :1827 Mean : 312.1   
## 3rd Qu.:0.0000 3rd Qu.:4.000 3rd Qu.:2300 3rd Qu.: 610.0   
## Max. :4.0000 Max. :5.000 Max. :9410 Max. :4820.0   
## yr\_built yr\_renovated street city   
## Min. :1900 Min. : 0.0 Length:4600 Length:4600   
## 1st Qu.:1951 1st Qu.: 0.0 Class :character Class :character   
## Median :1976 Median : 0.0 Mode :character Mode :character   
## Mean :1971 Mean : 808.6   
## 3rd Qu.:1997 3rd Qu.:1999.0   
## Max. :2014 Max. :2014.0   
## statezip country   
## Length:4600 Length:4600   
## Class :character Class :character   
## Mode :character Mode :character   
##   
##   
##

colnames(data)

## [1] "date" "price" "bedrooms" "bathrooms"   
## [5] "sqft\_living" "sqft\_lot" "floors" "waterfront"   
## [9] "view" "condition" "sqft\_above" "sqft\_basement"  
## [13] "yr\_built" "yr\_renovated" "street" "city"   
## [17] "statezip" "country"

str(data)

## 'data.frame': 4600 obs. of 18 variables:  
## $ date : chr "2014-05-02 00:00:00" "2014-05-02 00:00:00" "2014-05-02 00:00:00" "2014-05-02 00:00:00" ...  
## $ price : num 313000 2384000 342000 420000 550000 ...  
## $ bedrooms : num 3 5 3 3 4 2 2 4 3 4 ...  
## $ bathrooms : num 1.5 2.5 2 2.25 2.5 1 2 2.5 2.5 2 ...  
## $ sqft\_living : int 1340 3650 1930 2000 1940 880 1350 2710 2430 1520 ...  
## $ sqft\_lot : int 7912 9050 11947 8030 10500 6380 2560 35868 88426 6200 ...  
## $ floors : num 1.5 2 1 1 1 1 1 2 1 1.5 ...  
## $ waterfront : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ view : int 0 4 0 0 0 0 0 0 0 0 ...  
## $ condition : int 3 5 4 4 4 3 3 3 4 3 ...  
## $ sqft\_above : int 1340 3370 1930 1000 1140 880 1350 2710 1570 1520 ...  
## $ sqft\_basement: int 0 280 0 1000 800 0 0 0 860 0 ...  
## $ yr\_built : int 1955 1921 1966 1963 1976 1938 1976 1989 1985 1945 ...  
## $ yr\_renovated : int 2005 0 0 0 1992 1994 0 0 0 2010 ...  
## $ street : chr "18810 Densmore Ave N" "709 W Blaine St" "26206-26214 143rd Ave SE" "857 170th Pl NE" ...  
## $ city : chr "Shoreline" "Seattle" "Kent" "Bellevue" ...  
## $ statezip : chr "WA 98133" "WA 98119" "WA 98042" "WA 98008" ...  
## $ country : chr "USA" "USA" "USA" "USA" ...

#checking variables

length(unique(data$city)) #There are 44 unique city.

## [1] 44

length(unique(data$country)) #There is only 1 country which should mean nothing

## [1] 1

length(unique(data$statezip)) #77 different kinds of statezip

## [1] 77

length(unique(data$street)) #4525 that is too much

## [1] 4525

#From this result we are going to drop “street” and “country” variables because one country means no effect on prediction and 4525 country seems to be too much to include.

We will convert “city” and “state” into categorical variables.

new\_data = subset(data, select = -c(street, country, statezip, city) )  
head(new\_data)

## date price bedrooms bathrooms sqft\_living sqft\_lot floors  
## 1 2014-05-02 00:00:00 313000 3 1.50 1340 7912 1.5  
## 2 2014-05-02 00:00:00 2384000 5 2.50 3650 9050 2.0  
## 3 2014-05-02 00:00:00 342000 3 2.00 1930 11947 1.0  
## 4 2014-05-02 00:00:00 420000 3 2.25 2000 8030 1.0  
## 5 2014-05-02 00:00:00 550000 4 2.50 1940 10500 1.0  
## 6 2014-05-02 00:00:00 490000 2 1.00 880 6380 1.0  
## waterfront view condition sqft\_above sqft\_basement yr\_built yr\_renovated  
## 1 0 0 3 1340 0 1955 2005  
## 2 0 4 5 3370 280 1921 0  
## 3 0 0 4 1930 0 1966 0  
## 4 0 0 4 1000 1000 1963 0  
## 5 0 0 4 1140 800 1976 1992  
## 6 0 0 3 880 0 1938 1994

str(new\_data)

## 'data.frame': 4600 obs. of 14 variables:  
## $ date : chr "2014-05-02 00:00:00" "2014-05-02 00:00:00" "2014-05-02 00:00:00" "2014-05-02 00:00:00" ...  
## $ price : num 313000 2384000 342000 420000 550000 ...  
## $ bedrooms : num 3 5 3 3 4 2 2 4 3 4 ...  
## $ bathrooms : num 1.5 2.5 2 2.25 2.5 1 2 2.5 2.5 2 ...  
## $ sqft\_living : int 1340 3650 1930 2000 1940 880 1350 2710 2430 1520 ...  
## $ sqft\_lot : int 7912 9050 11947 8030 10500 6380 2560 35868 88426 6200 ...  
## $ floors : num 1.5 2 1 1 1 1 1 2 1 1.5 ...  
## $ waterfront : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ view : int 0 4 0 0 0 0 0 0 0 0 ...  
## $ condition : int 3 5 4 4 4 3 3 3 4 3 ...  
## $ sqft\_above : int 1340 3370 1930 1000 1140 880 1350 2710 1570 1520 ...  
## $ sqft\_basement: int 0 280 0 1000 800 0 0 0 860 0 ...  
## $ yr\_built : int 1955 1921 1966 1963 1976 1938 1976 1989 1985 1945 ...  
## $ yr\_renovated : int 2005 0 0 0 1992 1994 0 0 0 2010 ...

#I want to change the code here to look better.   
  
library("stringr")   
tail(new\_data$date, n=100)

## [1] "2014-06-17 00:00:00" "2014-06-17 00:00:00" "2014-06-17 00:00:00"  
## [4] "2014-06-17 00:00:00" "2014-06-17 00:00:00" "2014-06-17 00:00:00"  
## [7] "2014-06-18 00:00:00" "2014-06-18 00:00:00" "2014-06-18 00:00:00"  
## [10] "2014-06-18 00:00:00" "2014-06-18 00:00:00" "2014-06-19 00:00:00"  
## [13] "2014-06-19 00:00:00" "2014-06-19 00:00:00" "2014-06-19 00:00:00"  
## [16] "2014-06-19 00:00:00" "2014-06-19 00:00:00" "2014-06-19 00:00:00"  
## [19] "2014-06-20 00:00:00" "2014-06-20 00:00:00" "2014-06-20 00:00:00"  
## [22] "2014-06-20 00:00:00" "2014-06-22 00:00:00" "2014-06-23 00:00:00"  
## [25] "2014-06-23 00:00:00" "2014-06-23 00:00:00" "2014-06-23 00:00:00"  
## [28] "2014-06-23 00:00:00" "2014-06-24 00:00:00" "2014-06-24 00:00:00"  
## [31] "2014-06-24 00:00:00" "2014-06-24 00:00:00" "2014-06-24 00:00:00"  
## [34] "2014-06-24 00:00:00" "2014-06-24 00:00:00" "2014-06-24 00:00:00"  
## [37] "2014-06-24 00:00:00" "2014-06-24 00:00:00" "2014-06-24 00:00:00"  
## [40] "2014-06-25 00:00:00" "2014-06-25 00:00:00" "2014-06-25 00:00:00"  
## [43] "2014-06-25 00:00:00" "2014-06-25 00:00:00" "2014-06-26 00:00:00"  
## [46] "2014-06-26 00:00:00" "2014-06-26 00:00:00" "2014-06-26 00:00:00"  
## [49] "2014-06-26 00:00:00" "2014-06-26 00:00:00" "2014-06-26 00:00:00"  
## [52] "2014-06-26 00:00:00" "2014-06-26 00:00:00" "2014-06-27 00:00:00"  
## [55] "2014-06-27 00:00:00" "2014-06-27 00:00:00" "2014-06-27 00:00:00"  
## [58] "2014-06-27 00:00:00" "2014-06-28 00:00:00" "2014-06-29 00:00:00"  
## [61] "2014-06-30 00:00:00" "2014-06-30 00:00:00" "2014-06-30 00:00:00"  
## [64] "2014-07-01 00:00:00" "2014-07-01 00:00:00" "2014-07-01 00:00:00"  
## [67] "2014-07-01 00:00:00" "2014-07-02 00:00:00" "2014-07-02 00:00:00"  
## [70] "2014-07-02 00:00:00" "2014-07-02 00:00:00" "2014-07-02 00:00:00"  
## [73] "2014-07-02 00:00:00" "2014-07-02 00:00:00" "2014-07-02 00:00:00"  
## [76] "2014-07-02 00:00:00" "2014-07-02 00:00:00" "2014-07-03 00:00:00"  
## [79] "2014-07-05 00:00:00" "2014-07-06 00:00:00" "2014-07-07 00:00:00"  
## [82] "2014-07-07 00:00:00" "2014-07-07 00:00:00" "2014-07-07 00:00:00"  
## [85] "2014-07-07 00:00:00" "2014-07-07 00:00:00" "2014-07-07 00:00:00"  
## [88] "2014-07-08 00:00:00" "2014-07-08 00:00:00" "2014-07-08 00:00:00"  
## [91] "2014-07-08 00:00:00" "2014-07-08 00:00:00" "2014-07-08 00:00:00"  
## [94] "2014-07-08 00:00:00" "2014-07-09 00:00:00" "2014-07-09 00:00:00"  
## [97] "2014-07-09 00:00:00" "2014-07-09 00:00:00" "2014-07-10 00:00:00"  
## [100] "2014-07-10 00:00:00"

str\_count(new\_data$date, "2014")

## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [112] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [149] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [186] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [223] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [260] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [297] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [334] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [371] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [408] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [445] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [482] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [519] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [556] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [593] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [630] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [667] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [704] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [741] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [778] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [815] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [852] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [889] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [926] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [963] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1000] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1037] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1074] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1111] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1148] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1185] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1222] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1259] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1296] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1333] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1370] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1407] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1444] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1481] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1518] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1555] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1592] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1629] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1666] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1703] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1740] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1777] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1814] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1851] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1888] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1925] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1962] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [1999] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2036] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2073] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2110] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2147] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2184] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2221] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2258] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2295] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2332] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2369] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2406] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2443] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2480] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2517] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2554] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2591] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2628] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2665] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2702] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2739] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2776] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2813] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2850] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2887] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2924] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2961] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [2998] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3035] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3072] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3109] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3146] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3183] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3220] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3257] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3294] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3331] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3368] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3405] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3442] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3479] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3516] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3553] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3590] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3627] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3664] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3701] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3738] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3775] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3812] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3849] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3886] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3923] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3960] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [3997] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [4034] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [4071] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [4108] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [4145] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [4182] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [4219] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [4256] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [4293] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [4330] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [4367] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [4404] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [4441] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [4478] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [4515] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [4552] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [4589] 1 1 1 1 1 1 1 1 1 1 1 1

It looks like all the housings are from the year of 2014.Therefore we are just going to drop this column and make a new column that indicate the years of the house. To show that, we are going to subtract “2014” by “the years it was built”.

new\_data = subset(new\_data, select = -c(date))  
sum(is.na(new\_data))

## [1] 0

head(new\_data)

## price bedrooms bathrooms sqft\_living sqft\_lot floors waterfront view  
## 1 313000 3 1.50 1340 7912 1.5 0 0  
## 2 2384000 5 2.50 3650 9050 2.0 0 4  
## 3 342000 3 2.00 1930 11947 1.0 0 0  
## 4 420000 3 2.25 2000 8030 1.0 0 0  
## 5 550000 4 2.50 1940 10500 1.0 0 0  
## 6 490000 2 1.00 880 6380 1.0 0 0  
## condition sqft\_above sqft\_basement yr\_built yr\_renovated  
## 1 3 1340 0 1955 2005  
## 2 5 3370 280 1921 0  
## 3 4 1930 0 1966 0  
## 4 4 1000 1000 1963 0  
## 5 4 1140 800 1976 1992  
## 6 3 880 0 1938 1994

fit <- lm(price ~ bedrooms + bathrooms + sqft\_living + sqft\_lot + floors + waterfront + view + condition + sqft\_above + sqft\_basement + yr\_built + yr\_renovated, data = new\_data)  
summary(fit)

##   
## Call:  
## lm(formula = price ~ bedrooms + bathrooms + sqft\_living + sqft\_lot +   
## floors + waterfront + view + condition + sqft\_above + sqft\_basement +   
## yr\_built + yr\_renovated, data = new\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2149360 -128320 -17027 89256 26332889   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.584e+06 6.853e+05 6.689 2.51e-11 \*\*\*  
## bedrooms -5.804e+04 1.049e+04 -5.531 3.36e-08 \*\*\*  
## bathrooms 5.720e+04 1.701e+04 3.363 0.000777 \*\*\*  
## sqft\_living 2.318e+02 2.168e+01 10.690 < 2e-16 \*\*\*  
## sqft\_lot -6.912e-01 2.127e-01 -3.250 0.001162 \*\*   
## floors 3.981e+04 1.870e+04 2.129 0.033346 \*   
## waterfront 3.553e+05 9.378e+04 3.789 0.000153 \*\*\*  
## view 4.570e+04 1.097e+04 4.167 3.14e-05 \*\*\*  
## condition 3.184e+04 1.304e+04 2.441 0.014680 \*   
## sqft\_above 2.966e+01 2.160e+01 1.374 0.169632   
## sqft\_basement NA NA NA NA   
## yr\_built -2.378e+03 3.416e+02 -6.962 3.84e-12 \*\*\*  
## yr\_renovated 6.573e+00 8.634e+00 0.761 0.446560   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 499300 on 4588 degrees of freedom  
## Multiple R-squared: 0.2178, Adjusted R-squared: 0.2159   
## F-statistic: 116.1 on 11 and 4588 DF, p-value: < 2.2e-16

##Linear Regression with all variables look quite bad model since R squared is around 0.3. Let’s drop two categorical columns “statezip” and “city”

new\_fit <- lm(price ~ bedrooms + bathrooms + sqft\_living + sqft\_lot + floors + waterfront + view + condition + sqft\_above + sqft\_basement + yr\_built + yr\_renovated, data = new\_data)  
summary(new\_fit)

##   
## Call:  
## lm(formula = price ~ bedrooms + bathrooms + sqft\_living + sqft\_lot +   
## floors + waterfront + view + condition + sqft\_above + sqft\_basement +   
## yr\_built + yr\_renovated, data = new\_data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2149360 -128320 -17027 89256 26332889   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.584e+06 6.853e+05 6.689 2.51e-11 \*\*\*  
## bedrooms -5.804e+04 1.049e+04 -5.531 3.36e-08 \*\*\*  
## bathrooms 5.720e+04 1.701e+04 3.363 0.000777 \*\*\*  
## sqft\_living 2.318e+02 2.168e+01 10.690 < 2e-16 \*\*\*  
## sqft\_lot -6.912e-01 2.127e-01 -3.250 0.001162 \*\*   
## floors 3.981e+04 1.870e+04 2.129 0.033346 \*   
## waterfront 3.553e+05 9.378e+04 3.789 0.000153 \*\*\*  
## view 4.570e+04 1.097e+04 4.167 3.14e-05 \*\*\*  
## condition 3.184e+04 1.304e+04 2.441 0.014680 \*   
## sqft\_above 2.966e+01 2.160e+01 1.374 0.169632   
## sqft\_basement NA NA NA NA   
## yr\_built -2.378e+03 3.416e+02 -6.962 3.84e-12 \*\*\*  
## yr\_renovated 6.573e+00 8.634e+00 0.761 0.446560   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 499300 on 4588 degrees of freedom  
## Multiple R-squared: 0.2178, Adjusted R-squared: 0.2159   
## F-statistic: 116.1 on 11 and 4588 DF, p-value: < 2.2e-16

=