```
King County House Sales
Elena Kargopiltsev
1/17/2021
#1. Used the read.csv() function to read the data into R. Called the loaded data House. Made sure that the working directory was set to the correct
location for the data. If not, use the setwd() function to change the working directory to the right location. Used the attach() function to attach House
to the R search path.
 House=read.csv("kc house sales.csv")
 attach(House)
#2. Used the summary() function to obtain a numerical summary of the variables in the data set.
 summary(House)
 ##
         date
                           price
                                           bedrooms
                                                          bathrooms
 ##
    Length: 21613
                       Min. : 75000 Min. : 0.000 Min. :0.000
    ##
     Mode :character Median : 450000 Median : 3.000
                                                         Median :2.250
 ##
                       Mean : 540182 Mean : 3.371 Mean :2.115
 ##
                       3rd Qu.: 645000 3rd Qu.: 4.000 3rd Qu.:2.500
                       Max. :7700000 Max. :33.000 Max. :8.000
 ##
                   sqft_lot floors waterfront
 ##
      sqft_living
     Min. : 290 Min. : 520 Min. :1.000 Min. :0.000000
                    1st Qu.: 5040 1st Qu.:1.000 1st Qu.:0.000000
     1st Qu.: 1427
                    Median : 7618
 ##
     Median : 1910
                                     Median :1.500 Median :0.000000
     Mean : 2080
                    Mean : 15107 Mean :1.494 Mean :0.007542
 ##
     3rd Qu.: 2550
                    3rd Qu.: 10688 3rd Qu.:2.000 3rd Qu.:0.000000
     Max. :13540 Max. :1651359 Max. :3.500 Max. :1.000000
 ##
                    condition grade sqft_above
 ##
         view
    Min. :0.0000 Min. :1.000 Min. : 1.000 Min. : 290
 ##
 ##
     Median :0.0000 Median :3.000 Median : 7.000 Median :1560
     Mean :0.2343 Mean :3.409 Mean : 7.657 Mean :1788
 ##
     3rd Qu.:0.0000 3rd Qu.:4.000 3rd Qu.: 8.000 3rd Qu.:2210
    Max. :4.0000 Max. :5.000 Max. :13.000 Max. :9410
                    yr_built yr_renovated
    sqft_basement
                                                    zipcode
 ##
 ##
    Min. : 0.0 Min. :1900 Min. : 0.0 Min. :98001
     1st Qu.: 0.0 1st Qu.:1951 1st Qu.: 0.0 1st Qu.:98033
 ##
    Median : 0.0 Median :1975 Median : 0.0
                                                    Median :98065
     Mean : 291.5 Mean :1971 Mean : 84.4
                                                    Mean :98078
 ##
    3rd Qu.: 560.0 3rd Qu.:1997 3rd Qu.: 0.0
                                                    3rd Qu.:98118
    Max. :4820.0 Max. :2015 Max. :2015.0
 ##
                                                    Max. :98199
 ##
         lat
                    long
                                    sqft_living15
                                                    sqft_lot15
    Min. :47.16 Min. :-122.5 Min. : 399 Min. : 651
 ##
     Median :47.57
                    Median :-122.2 Median :1840
                                                   Median : 7620
    Mean :47.56
                    Mean :-122.2 Mean :1987
                                                  Mean : 12768
 ## 3rd Qu.:47.68
                    3rd Qu.:-122.1 3rd Qu.:2360 3rd Qu.: 10083
 ## Max. :47.78 Max. :-121.3 Max. :6210 Max. :871200
#3. Among the variables, the date was treated as a categorical variable in the numerical summary. Converted it to dates using the as. Date()
function. In the function, the upper case Y means the first 4 digits represent the year, while lower case m and d mean the next 2 digits each
represent month and day, respectively.
 date=as.Date(date, format="%Y%m%d")
#4. Waterfront, view, and zipcode variables were categorical; however, they were treated as numerical variables in the numerical summary.
Converted them to categorical variables using the as.factor() function
 view=as.factor(view)
 waterfront=as.factor(waterfront)
 zipcode=as.factor(zipcode)
#5.Obtained a numerical summary of the data set one more time using summary(House).
 summary(House)
                           price
         date
                                           bedrooms
                                                           bathrooms
     Length: 21613
                             : 75000
                                        Min. : 0.000
                                                         Min.
                                                               :0.000
                       1st Qu.: 321950
     Class :character
                                        1st Qu.: 3.000
 ##
                                                         1st Qu.:1.750
     Mode :character
                       Median : 450000
                                        Median : 3.000
                                                         Median :2.250
 ##
                       Mean : 540182
                                        Mean : 3.371
                                                         Mean :2.115
 ##
                       3rd Qu.: 645000
                                        3rd Qu.: 4.000
                                                         3rd Qu.:2.500
 ##
                       Max. :7700000
                                        Max. :33.000
                                                         Max.
                                                               :8.000
 ##
      sqft_living
                       sqft_lot
                                         floors
                                                       waterfront
 ##
     Min. : 290
                                520
                                     Min. :1.000
                                                     Min. :0.000000
                    Min. :
     1st Qu.: 1427
                    1st Qu.:
                              5040
                                     1st Qu.:1.000
                                                     1st Qu.:0.000000
     Median : 1910
                    Median :
                              7618
                                      Median :1.500
                                                     Median :0.000000
 ##
     Mean
          : 2080
                    Mean : 15107
                                      Mean :1.494
                                                     Mean :0.007542
 ##
     3rd Qu.: 2550
                    3rd Qu.: 10688
                                     3rd Qu.:2.000
                                                     3rd Qu.:0.000000
 ##
     Max.
           :13540
                    Max. :1651359
                                     Max. :3.500
                                                     Max. :1.000000
 ##
          view
                       condition
                                        grade
                                                       sqft_above
 ##
     Min.
            :0.0000
                     Min. :1.000
                                     Min. : 1.000
                                                     Min. : 290
 ##
     1st Qu.:0.0000
                     1st Qu.:3.000
                                     1st Qu.: 7.000
                                                     1st Qu.:1190
 ##
     Median :0.0000
                     Median :3.000
                                     Median : 7.000
                                                     Median:1560
 ##
            :0.2343
                     Mean :3.409
                                     Mean : 7.657
                                                     Mean :1788
     3rd Qu.:0.0000
                                                     3rd Qu.:2210
 ##
                     3rd Qu.:4.000
                                     3rd Qu.: 8.000
            :4.0000
                     Max. :5.000
                                     Max. :13.000
                                                     Max.
                                                            :9410
     sqft_basement
                        yr_built
 ##
                                     yr_renovated
                                                       zipcode
 ##
               0.0
                     Min. :1900
                                    Min. : 0.0
                                                          :98001
     Min.
                                                    Min.
     1st Qu.:
               0.0
                     1st Qu.:1951
                                    1st Qu.: 0.0
                                                    1st Qu.:98033
                     Median :1975
                                    Median :
                                              0.0
                                                    Median :98065
 ##
     Median :
               0.0
 ##
     Mean
          : 291.5
                     Mean :1971
                                   Mean : 84.4
                                                    Mean :98078
     3rd Qu.: 560.0
                     3rd Qu.:1997
                                    3rd Qu.: 0.0
                                                    3rd Qu.:98118
            :4820.0
                     Max. :2015
                                   Max. :2015.0
                                                          :98199
 ##
     Max.
                                                    Max.
 ##
          lat
                         long
                                     sqft_living15
                                                     sqft_lot15
 ##
                                           : 399
            :47.16
                    Min. :-122.5 Min.
                                                   Min.
     1st Qu.:47.47
                    1st Qu.:-122.3
                                     1st Qu.:1490
                                                   1st Qu.: 5100
 ##
     Median :47.57
                    Median :-122.2
                                                   Median: 7620
 ##
                                     Median :1840
           :47.56
                    Mean :-122.2
                                     Mean
                                           :1987
                                                   Mean
                                                         : 12768
 ##
     3rd Qu.:47.68
                    3rd Qu.:-122.1
                                     3rd Qu.:2360
                                                   3rd Qu.: 10083
            :47.78
                          :-121.3
                                     Max.
                                           :6210
                                                   Max.
                                                          :871200
As a result, all four variables are still considered numerical.
#6.Obtained a numerical summary of the four variables created in the global environment in question 3 and 4 using summary(data.frame(date,
view, waterfront, zipcode)).
 summary(data.frame(date, view, waterfront, zipcode))
 ##
          date
                         view
                                   waterfront
                                                zipcode
 ##
           :2014-05-02
                         0:19489
                                   0:21450
                                             98103
                                                   : 602
     Min.
     1st Qu.:2014-07-22
                         1: 332
                                             98038
                                                       590
     Median :2014-10-16
                            963
                                                    :
 ##
                         2:
                                             98115
                                                       583
 ##
            :2014-10-29
                         3: 510
                                             98052
                                                   : 574
     3rd Qu.:2015-02-17
                                             98117
                                                    : 553
 ##
     Max.
            :2015-05-27
                                             98042
                                                   : 548
 ##
                                             (Other):18163
At this time variables treated correctly because I used
"data.frames".
#7. Eliminated the four variables in House and add date, view, waterfront, and zip code in the global environment to House. Obtained numerical
summary of the variables in-house.
 House=data.frame(House[,-c(1,8:9,16)], date, view, waterfront, zipcode)
 summary(House)
 ##
         price
                         bedrooms
                                        bathrooms
                                                       sqft_living
 ## Min. : 75000 Min. : 0.000 Min. :0.000 Min. : 290
    1st Qu.: 1427
     Median : 450000
                                                      Median : 1910
                      Median : 3.000 Median :2.250
     Mean : 540182 Mean : 3.371 Mean : 2.115 Mean : 2080
 ##
     3rd Qu.: 645000 3rd Qu.: 4.000 3rd Qu.:2.500 3rd Qu.: 2550
     Max. :7700000 Max. :33.000 Max. :8.000 Max. :13540
 ##
 ##
                                       condition
        sqft_lot
                       floors
                                                         grade
 ##
     Min. : 520 Min. :1.000 Min. :1.000 Min. : 1.000
                                                    1st Qu.: 7.000
 ##
     1st Qu.:
               5040
                      1st Qu.:1.000
                                     1st Qu.:3.000
               7618 Median :1.500
                                     Median :3.000
                                                    Median : 7.000
     Mean : 15107
                      Mean :1.494
                                     Mean :3.409
                                                    Mean : 7.657
 ##
                                     3rd Qu.:4.000 3rd Qu.: 8.000
     3rd Qu.: 10688 3rd Qu.:2.000
 ##
     Max. :1651359 Max. :3.500 Max. :5.000 Max. :13.000
 ##
                                      yr_built
 ##
       sqft_above sqft_basement
                                                   yr_renovated
                   Min. : 0.0 Min. :1900
     Min. : 290
 ##
     1st Qu.:
                                                            0.0
     Median :1560
 ##
                   Median: 0.0 Median:1975
                                                  Median : 0.0
     Mean :1788
                  Mean : 291.5 Mean :1971
                                                  Mean : 84.4
     3rd Qu.:2210 3rd Qu.: 560.0 3rd Qu.:1997
                                                  3rd Qu.: 0.0
 ##
           :9410 Max. :4820.0 Max. :2015 Max. :2015.0
 ##
     Max.
 ##
         lat
 ##
                         long
                                     sqft_living15
                                                   sqft_lot15
     Min. :47.16 Min. :-122.5 Min. : 399 Min. : 651
 ##
     1st Qu.:47.47    1st Qu.:-122.3    1st Qu.:1490    1st Qu.: 5100
     Median :47.57 Median :-122.2 Median :1840 Median : 7620
     Mean :47.56
                    Mean :-122.2 Mean :1987
                                                   Mean : 12768
 ##
     3rd Qu.:47.68 3rd Qu.:-122.1 3rd Qu.:2360
                                                   3rd Qu.: 10083
          :47.78 Max. :-121.3 Max. :6210 Max. :871200
    Max.
 ##
 ##
          date
                         view
                                   waterfront zipcode
     Min. :2014-05-02 0:19489 0:21450 98103 : 602
     1st Qu.:2014-07-22 1: 332 1: 163 98038 : 590
 ##
     Median: 2014-10-16 2: 963
Mean: 2014-10-29 3: 510
3rd Qu.: 2015-02-17 4: 319
 ##
                                             98115 : 583
                                             98052 : 574
     3rd Qu.:2015-02-17 4: 319
 ##
                                             98117 : 553
 ##
     Max. :2015-05-27
                                             98042 : 548
 ##
                                             (Other):18163
#8. Used the unique() function to produce different levels of zipcode. To analyze how many levels it has.
 unique(zipcode)
 ## [1] 98178 98125 98028 98136 98074 98053 98003 98198 98146 98038 98007 98115
 ## [13] 98107 98126 98019 98103 98002 98133 98040 98092 98030 98119 98112 98052
 ## [25] 98027 98117 98058 98001 98056 98166 98023 98070 98148 98105 98042 98008
 ## [37] 98059 98122 98144 98004 98005 98034 98075 98116 98010 98118 98199 98032
 ## [49] 98045 98102 98077 98108 98168 98177 98065 98029 98006 98109 98022 98033
 ## [61] 98155 98024 98011 98031 98106 98072 98188 98014 98055 98039
 ## 70 Levels: 98001 98002 98003 98004 98005 98006 98007 98008 98010 ... 98199
It has 70 zipcode levels.
#9.Used the pairs() function to obtain a scatter-plot matrix of price, bedrooms, and bathrooms
 pairs(House[,1:3])
                               0 5 10 15 20 25 30
               price
                                    bedrooms
                                                           bathrooms
      0e+00 2e+06 4e+06 6e+06 8e+06
#10.Used the table() function to produce a frequency table of bedrooms. Analysing any unusual observations.
 table(bedrooms)
 ## bedrooms
      0 1 2 3 4 5 6
 ## 13 199 2760 9824 6882 1601 272 38 13
There are 33 bedrooms in one house.
#11.One house has 33 bedrooms. Printing the observations with more than 30 bedrooms
 House[bedrooms>30,]
           price bedrooms bathrooms sqft_living sqft_lot floors condition grade
 ## 15871 640000
                  33 1.75
                                     1620 6000 1
 ##
         sqft_above sqft_basement yr_built yr_renovated lat
                                                                  long
 ## 15871 1040 580 1947 0 47.6878 -122.331
         ## 15871 1330 4700 2014-06-25 0
                                                         0 98103
#12. It is obvious that the above observation (observation #15871) was incorrectly recorded since it had 33 bedrooms while its living space was
only 1620 sf. Changing the number of bedrooms of this observation to 3 using the following function:
 House[15871, 2]=3
 detach(House)
 attach(House)
 ## The following objects are masked _by_ .GlobalEnv:
 ##
        date, view, waterfront, zipcode
#13. Used the pairs() function to obtain a scatter-plot matrix of price, bedrooms, and bathrooms one more time. Discovered a relationship between
the three variables.
 pairs(House[,1:3])
                                  2
               price
                                                           0 0 0
00000000 0 0
00000000000 00
                                                         0 000000000000000 0 0
                                                                                   ## The price of a house increases
                                    bedrooms
                                                        000 000000000000000000
                                                        000000000
                                                           bathrooms
                                                                     6
      0e+00 2e+06 4e+06
                     6e+06 8e+06
with the number of bedrooms and bathrooms.
#14.Used the plot() function to produce side-by-side box plots of price for two waterfront categories and five categories of view.
 plot(waterfront, price)
     8e+06
                           0
                                                           0
                            0
     6e+06
                                                           8
     4e+06
     0e+00
                           0
                                           X
 plot(view, price)
     8e+06
                                                        0
                                           0
     90+99
     4e+06
     2e+06
     0e+00
                                           2
                                                        3
                  0
                              1
                                           X
#15. Used the hist() function to produce histograms of price, sqft_living, and sqft_lot
 hist(price)
                                 Histogram of price
     10000
Frequency
     0009
     2000
     0
         0e+00
                         2e+06
                                         4e+06
                                                         6e+06
                                                                         8e+06
                                         price
 hist(price, breaks=30, col=2)
                                 Histogram of price
     8000
     0009
Frequency
     4000
     2000
     0
         0e+00
                          2e+06
                                                          6e+06
                                          4e+06
                                                                          8e+06
                                          price
 hist(sqft_living,col="green")
                              Histogram of sqft_living
     10000
     8000
     0009
Frequency
     4000
     2000
     0
                   2000
                                     6000
                                              8000
                                                      10000
           0
                            4000
                                                                12000
                                                                         14000
                                       sqft_living
 hist(sqft_lot)
                                Histogram of sqft_lot
     20000
     15000
     10000
     5000
     0
           0
                            500000
                                             1000000
                                                                1500000
                                        sqft_lot
 hist(sqft_lot[sqft_lot<30000], col="gray")</pre>
                      Histogram of sqft_lot[sqft_lot < 30000]
     4000
     3000
Frequency
     2000
                                                                                   #16. Used the command
     1000
     0
           0
                    5000
                              10000
                                         15000
                                                   20000
                                                              25000
                                                                         30000
                                 sqft_lot[sqft_lot < 30000]
par(mfrow=c(2, 2)) or par(mfcol=c(2, 2)) to divide the plots window into four regions and repeat questions 14 and 15. You can modify the
arguments to these functions to divide the window in a different way. To return to the original setting, Used the command par(mfrow=c(1, 1)).
 par(mfrow=c(2,2))
 plot(waterfront, price)
 plot(view, price)
 hist(price)
 hist(price, breaks=30, col=2)
                                              ee-
    0e+00
                                              0e+00
                      X
                                                               X
                                                       Histogram of price
              Histogram of price
                                              8000
Frequency
    0009
             2e+06
       0e+00
                    4e+06
                           6e+06
                                                0e+00
                                                       2e+06
                                                              4e+06
                                                                     6e+06 8e+06
                     price
                                                              price
 hist(sqft_living,col="green")
 hist(sqft_lot)
 hist(sqft_lot[sqft_lot<30000],col="gray")</pre>
                                                      Histogram of sqft_lot
           Histogram of sqft_living
    10000
Frequency
                                          Frequency
                                              10000
         0 2000
                  6000
                          10000
                                  14000
                                                        500000
                                                               1000000 1500000
                   sqft_living
                                                             sqft_lot
    Histogram of sqft_lot[sqft_lot < 30000]
Frequency
         0 5000
                    15000
                             25000
             sqft_lot[sqft_lot < 30000]
#17.Used the sapply() function to obtain the range of all quantitative variables and call the output ranges
 ranges=sapply(House[,1:16], range)
#18.Outputting the entire matrix of ranges. Printed the largest number of bathrooms in House data using the indies of ranges
 ranges
           price bedrooms bathrooms sqft_living sqft_lot floors condition grade
 ## [1,]
           75000
                                                   520
                                                         1.0
         sqft_above sqft_basement yr_built yr_renovated
                                                        lat
 ## [1,]
               290
                               0
                                     1900
                                                    0 47.1559 -122.519
                                                 2015 47.7776 -121.315
 ## [2,]
                                     2015
               9410
                            4820
         sqft_living15 sqft_lot15
 ## [1,]
                  399
                             651
 ## [2,]
                 6210
                          871200
 ranges[2,]
 ##
                      bedrooms
                                              sqft_living
                                                               sqft_lot
            price
                                   bathrooms
                       11.0000
                                     8.0000
     7700000.0000
                                               13540.0000 1651359.0000
                     condition
                                               sqft_above sqft_basement
 ##
           floors
                                      grade
           3.5000
                        5.0000
                                     13.0000
                                                9410.0000
 ##
                                                              4820.0000
         yr_built yr_renovated
 ##
                                        lat
                                                     long sqft_living15
                                     47.7776
        2015.0000
                     2015.0000
                                                              6210.0000
 ##
                                                -121.3150
 ##
       sqft_lot15
      871200.0000
#19. Created a new categorical variable, called a mansion, by binning the sqft_living15 variable. All houses are divided into two groups based on
whether or not the living space in 2015 exceeds 5,000 SF.
 mansion=rep("No", nrow(House))
```

mansion[sqft\_living15>5000]="Yes"

#20. Founded how many mansions were sold during the period using the summary() function.

#21. Used the plot() function to produce side-by-side box-plots of price versus mansion.

No

House.wf=subset(House, waterfront==1)

House.old=subset(House,yr\_built<=1920)</pre>

mean(House.wf\$price)

number of houses in House.old.

## [1] 1662524

dim(House.old)

## [1] 1549 20

Yes

#22.Used the subset() function to create a subset, called House.wf, containing waterfront houses only. Finded the average price of the waterfront

#23.Used the subset() function to create a subset, called House.old, containing houses at least 100 years old. Used the dim() function to find the

mansion=as.factor(mansion)

Yes

22

summary(mansion)

No

plot(mansion, price)

## 21591

6e+06

2e+06