# 5_2_Exercise.R

79bar

2022-10-02

```
# Assignment: ASSIGNMENT 5
# Name: Jean, Barbara
# Date: 2022-10-02


## Load the ggplot2 package

library("readxl")
library(dplyr)
```

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

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

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

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --

## v ggplot2 3.3.6     v purrr   0.3.4
## v tibble  3.1.8     v stringr 1.4.1
## v tidyr   1.2.1     v forcats 0.5.2
## v readr   2.1.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggplot2)

setwd("C:/Users/79bar/dsc520")

housing <- read_excel("C:/Users/79bar/dsc520/data/week-6-housing.xlsx")
```

```
## Replace Blanks in Column Names with gsub()
names(housing) <- gsub(" ", "_", names(housing))
## Convert colnames to lowercase case
names(housing) <- tolower(names(housing))

##Using the dplyr package, use the 6 different operations to analyze/transform the data
##GroupBy, Summarize, Mutate, Filter, Select, and Arrange

housing%>%
  select(sale_date,sale_price,addr_full,square_feet_total_living,sq_ft_lot,bedrooms,
         bath_full_count,year_built)%>%
  filter(year_built>=2000)%>%
  arrange(year_built)
```

```
## # A tibble: 6,321 x 8
##    sale_date           sale_pr~1 addr_~2 squar~3 sq_ft~4 bedro~5 bath_~6 year_~7
##    <dttm>                  <dbl> <chr>     <dbl>   <dbl>   <dbl>   <dbl>   <dbl>
##  1 2006-02-09 00:00:00    647500 9206 1~    3620    4669       5       3    2000
##  2 2006-02-15 00:00:00   1390000 19656 ~    3280  225640       3       2    2000
##  3 2006-02-24 00:00:00    532000 10119 ~    2760    5160       4       2    2000
##  4 2006-06-06 00:00:00   1650000 2005 2~    5640   36172       4       3    2000
##  5 2006-07-19 00:00:00    804000 4521 2~     820   44236       0       1    2000
##  6 2006-09-12 00:00:00    777000 7228 1~    3160    4676       3       2    2000
##  7 2006-09-18 00:00:00   1230000 2208 2~    5340   16821       6       3    2000
##  8 2006-12-04 00:00:00    781000 14819 ~    3200    5651       3       2    2000
##  9 2007-03-29 00:00:00    999900 3343 W~    3260   27440       4       2    2000
## 10 2007-04-11 00:00:00    621000 2484 1~    2170    3840       3       2    2000
## # ... with 6,311 more rows, and abbreviated variable names 1: sale_price,
## #   2: addr_full, 3: square_feet_total_living, 4: sq_ft_lot, 5: bedrooms,
## #   6: bath_full_count, 7: year_built
```

```
housing%>%
  select(sale_price,sq_ft_lot)%>%
mutate(price_per_ftsq= sale_price/sq_ft_lot)
```

```
## # A tibble: 12,865 x 3
##    sale_price sq_ft_lot price_per_ftsq
##         <dbl>     <dbl>          <dbl>
##  1     698000      6635         105.
##  2     649990      5570         117.
##  3     572500      8444          67.8
##  4     420000      9600          43.8
##  5     369900      7526          49.1
##  6     184667      7280          25.4
##  7    1050000     97574          10.8
##  8     875000     30649          28.5
##  9     660000     42688          15.5
## 10     650000     94889           6.85
## # ... with 12,855 more rows
```

```
housing%>%
  select(sale_date,sale_price,addr_full,square_feet_total_living,sq_ft_lot,bedrooms,
```

```
        bath_full_count,year_built)%>%
  group_by(bedrooms)%>%
  summarise(avg_sq_ft= mean(sq_ft_lot), count=n(),na.rd=TRUE)
```

```
## # A tibble: 12 x 4
##    bedrooms avg_sq_ft count na.rd
##       <dbl>     <dbl> <int> <lgl>
##  1        0    56001.    19 TRUE
##  2        1    99499.    33 TRUE
##  3        2    14126.  1658 TRUE
##  4        3    21207.  4493 TRUE
##  5        4    24079.  5515 TRUE
##  6        5    24000.  1047 TRUE
##  7        6    39457.    83 TRUE
##  8        7   111931.    11 TRUE
##  9        8   219106      2 TRUE
## 10        9     9462      2 TRUE
## 11       10    17328      1 TRUE
## 12       11    13220      1 TRUE
```

```
##Using the purrr package - perform 2 functions on your dataset.


keep(housing$sale_price, ~ .x <50000)
```

```
##  [1] 31272 32000  5000  1000 48475 20000 46031 40191  7276 39000  5896 12500
## [13] 12500 10570 48740 41500  1500 20000 32000 20713 20146 45000 29537 47500
## [25]  1500   998   873   873  6000  1070   698   698 40000 38201  4000  4059
## [37] 40000 20000 42182  2031 35000 15000 14000  2500  7000  8000  8000  5150
## [49]  5150 18000 37800
```

```
discard(housing$sale_price, ~ .x>50000)
```

```
##  [1] 31272 32000  5000  1000 48475 20000 46031 40191  7276 39000  5896 12500
## [13] 12500 10570 48740 41500  1500 20000 32000 20713 20146 45000 29537 47500
## [25]  1500   998   873   873  6000 50000  1070   698   698 40000 38201  4000
## [37]  4059 50000 40000 20000 50000 42182  2031 35000 15000 14000  2500  7000
## [49]  8000  8000  5150  5150 18000 37800
```

```
##Use the cbind and rbind function on your dataset

## rbind function
nrow(housing)
```

```
## [1] 12865
```

```
housing_row1<- housing[1:10,]
housing_row2 <- housing[11:20,]
housing_rbind<-rbind(housing_row1,housing_row2)

## cbind function
ncol(housing)
```

```
## [1] 24
```

```
housing_col1<- housing[,1:6,]
housing_col2 <- housing[,7:12,]
housing_cbind<-cbind(housing_col1,housing_col2)


## Split a string, then concatenate the results back together

##Split a string
addr_split<- unlist(strsplit(housing$addr_full[1:10],split=" "))
addr_split
```

```
##  [1] "17021" "NE"    "113TH" "CT"    "11927" "178TH" "PL"    "NE"    "13315"
## [10] "174TH" "AVE"   "NE"    "3303"  "178TH" "AVE"   "NE"    "16126" "NE"
## [19] "108TH" "CT"    "8101"  "229TH" "DR"    "NE"    "21634" "NE"    "87TH"
## [28] "PL"    "21404" "NE"    "67TH"  "ST"    "7525"  "238TH" "AVE"   "NE"
## [37] "17703" "NE"    "26TH"  "ST"
```

```
##concatenate the results back together
sapply(list(addr_split),paste,collapse=" ")
```

```
## [1] "17021 NE 113TH CT 11927 178TH PL NE 13315 174TH AVE NE 3303 178TH AVE NE 16126 NE 108TH CT 8101
```