# Week 5 Assignment

Ruben Brionez Jr

2024-01-14

```
# Assignment: Week 5 Assignment
# Name: Ruben Brionez Jr
# Date: 01/14/2024
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.3.2
```

```
##
## 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(purrr)
```

```
## Warning: package 'purrr' was built under R version 4.3.2
```

```
library(tibble)
```

```
## Warning: package 'tibble' was built under R version 4.3.2
```

```
library(xml2)
```

```
## Warning: package 'xml2' was built under R version 4.3.2
```

```
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 4.3.2
```

```
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 4.3.2
```

```r
# Read the excel file and save to a variable
housing_df <- read_xlsx(path = "week-6-housing.xlsx")
# Review the data frame and types
str(housing_df)
```

```
## tibble [12,865 x 24] (S3: tbl_df/tbl/data.frame)
##  $ Sale Date               : POSIXct[1:12865], format: "2006-01-03" "2006-01-03" ...
##  $ Sale Price              : num [1:12865] 698000 649990 572500 420000 369900 ...
##  $ sale_reason             : num [1:12865] 1 1 1 1 1 1 1 1 1 1 ...
##  $ sale_instrument         : num [1:12865] 3 3 3 3 3 15 3 3 3 3 ...
##  $ sale_warning            : chr [1:12865] NA NA NA NA ...
##  $ sitetype                : chr [1:12865] "R1" "R1" "R1" "R1" ...
##  $ addr_full               : chr [1:12865] "17021 NE 113TH CT" "11927 178TH PL NE" "13315 174TH AVE N
##  $ zip5                    : num [1:12865] 98052 98052 98052 98052 98052 ...
##  $ ctyname                 : chr [1:12865] "REDMOND" "REDMOND" NA "REDMOND" ...
##  $ postalctyn              : chr [1:12865] "REDMOND" "REDMOND" "REDMOND" "REDMOND" ...
##  $ lon                     : num [1:12865] -122 -122 -122 -122 -122 ...
##  $ lat                     : num [1:12865] 47.7 47.7 47.7 47.6 47.7 ...
##  $ building_grade          : num [1:12865] 9 9 8 8 7 7 10 10 9 8 ...
##  $ square_feet_total_living: num [1:12865] 2810 2880 2770 1620 1440 4160 3960 3720 4160 2760 ...
##  $ bedrooms                : num [1:12865] 4 4 4 3 3 4 5 4 4 4 ...
##  $ bath_full_count         : num [1:12865] 2 2 1 1 1 2 3 2 2 1 ...
##  $ bath_half_count         : num [1:12865] 1 0 1 0 0 1 0 1 1 0 ...
##  $ bath_3qtr_count         : num [1:12865] 0 1 1 1 1 1 1 0 1 1 ...
##  $ year_built              : num [1:12865] 2003 2006 1987 1968 1980 ...
##  $ year_renovated          : num [1:12865] 0 0 0 0 0 0 0 0 0 0 ...
##  $ current_zoning          : chr [1:12865] "R4" "R4" "R6" "R4" ...
##  $ sq_ft_lot               : num [1:12865] 6635 5570 8444 9600 7526 ...
##  $ prop_type               : chr [1:12865] "R" "R" "R" "R" ...
##  $ present_use             : num [1:12865] 2 2 2 2 2 2 2 2 2 2 ...
```

```r
# Renaming 'Sale Price' to remove quotes
colnames(housing_df)[colnames(housing_df) == 'Sale Price'] <- 'sale_price'
# Verifying name change
str(housing_df)
```

```
## tibble [12,865 x 24] (S3: tbl_df/tbl/data.frame)
##  $ Sale Date               : POSIXct[1:12865], format: "2006-01-03" "2006-01-03" ...
##  $ sale_price              : num [1:12865] 698000 649990 572500 420000 369900 ...
##  $ sale_reason             : num [1:12865] 1 1 1 1 1 1 1 1 1 1 ...
##  $ sale_instrument         : num [1:12865] 3 3 3 3 3 15 3 3 3 3 ...
##  $ sale_warning            : chr [1:12865] NA NA NA NA ...
##  $ sitetype                : chr [1:12865] "R1" "R1" "R1" "R1" ...
##  $ addr_full               : chr [1:12865] "17021 NE 113TH CT" "11927 178TH PL NE" "13315 174TH AVE N
##  $ zip5                    : num [1:12865] 98052 98052 98052 98052 98052 ...
##  $ ctyname                 : chr [1:12865] "REDMOND" "REDMOND" NA "REDMOND" ...
##  $ postalctyn              : chr [1:12865] "REDMOND" "REDMOND" "REDMOND" "REDMOND" ...
##  $ lon                     : num [1:12865] -122 -122 -122 -122 -122 ...
##  $ lat                     : num [1:12865] 47.7 47.7 47.7 47.6 47.7 ...
##  $ building_grade          : num [1:12865] 9 9 8 8 7 7 10 10 9 8 ...
##  $ square_feet_total_living: num [1:12865] 2810 2880 2770 1620 1440 4160 3960 3720 4160 2760 ...
##  $ bedrooms                : num [1:12865] 4 4 4 3 3 4 5 4 4 4 ...
##  $ bath_full_count         : num [1:12865] 2 2 1 1 1 2 3 2 2 1 ...
```

```
##  $ bath_half_count         : num [1:12865] 1 0 1 0 0 1 0 1 1 0 ...
##  $ bath_3qtr_count         : num [1:12865] 0 1 1 1 1 1 1 0 1 1 ...
##  $ year_built              : num [1:12865] 2003 2006 1987 1968 1980 ...
##  $ year_renovated          : num [1:12865] 0 0 0 0 0 0 0 0 0 0 ...
##  $ current_zoning          : chr [1:12865] "R4" "R4" "R6" "R4" ...
##  $ sq_ft_lot               : num [1:12865] 6635 5570 8444 9600 7526 ...
##  $ prop_type               : chr [1:12865] "R" "R" "R" "R" ...
##  $ present_use             : num [1:12865] 2 2 2 2 2 2 2 2 2 2 ...
```

```
# Using SELECT square_feet_total_living column
housing_df %>% select(square_feet_total_living)
```

```
## # A tibble: 12,865 x 1
##    square_feet_total_living
##                       <dbl>
##  1                     2810
##  2                     2880
##  3                     2770
##  4                     1620
##  5                     1440
##  6                     4160
##  7                     3960
##  8                     3720
##  9                     4160
## 10                     2760
## # i 12,855 more rows
```

```
# Using SUMMARIZE to find the average sale price
housing_df %>% summarize(AvgPrice=mean(sale_price))
```

```
## # A tibble: 1 x 1
##   AvgPrice
##      <dbl>
## 1  660738.
```

```
# Using FILTER and then SELECT to specify results
housing_df %>% filter(square_feet_total_living >= 2000) %>% select(
  sale_price, addr_full, square_feet_total_living)
```

```
## # A tibble: 8,644 x 3
##    sale_price addr_full          square_feet_total_living
##         <dbl> <chr>                                <dbl>
##  1     698000 17021 NE 113TH CT                     2810
##  2     649990 11927 178TH PL NE                     2880
##  3     572500 13315 174TH AVE NE                    2770
##  4     184667 8101 229TH DR NE                      4160
##  5    1050000 21634 NE 87TH PL                      3960
##  6     875000 21404 NE 67TH ST                      3720
##  7     660000 7525 238TH AVE NE                     4160
##  8     650000 17703 NE 26TH ST                      2760
##  9     599950 14924 NE 74TH CT                      2180
## 10     526787 7858 148TH CT NE                      2480
## # i 8,634 more rows
```

```r
# Chaining Pipes to refine results
housing_df %>% select(sale_price, addr_full,
  ctyname, square_feet_total_living) %>% filter(square_feet_total_living < 1500)
```

```
## # A tibble: 1,528 x 4
##    sale_price addr_full               ctyname square_feet_total_living
##         <dbl> <chr>                   <chr>                      <dbl>
##  1     369900 16126 NE 108TH CT       REDMOND                     1440
##  2     399900 24307 NE VINE MAPLE WAY <NA>                        1350
##  3     350000 6028 215TH AVE NE       <NA>                        1420
##  4     443509 24253 NE VINE MAPLE WAY <NA>                        1350
##  5     362000 13706 NE 76TH PL        REDMOND                     1250
##  6     345000 8607 139TH AVE NE       REDMOND                     1280
##  7     435000 7428 153RD CT NE        REDMOND                     1460
##  8     390750 22315 NE 98TH ST        <NA>                        1440
##  9     405000 17231 NE 133RD PL       <NA>                        1320
## 10     275000 18122 NE 91ST CT        REDMOND                      920
## # i 1,518 more rows
```

```r
# Using MUTATE to find price per sq ft
housing_df %>% select(sale_price, square_feet_total_living) %>%
  mutate(price_sq_ft=sale_price/square_feet_total_living)
```

```
## # A tibble: 12,865 x 3
##    sale_price square_feet_total_living price_sq_ft
##         <dbl>                    <dbl>       <dbl>
##  1     698000                     2810       248.
##  2     649990                     2880       226.
##  3     572500                     2770       207.
##  4     420000                     1620       259.
##  5     369900                     1440       257.
##  6     184667                     4160        44.4
##  7    1050000                     3960       265.
##  8     875000                     3720       235.
##  9     660000                     4160       159.
## 10     650000                     2760       236.
## # i 12,855 more rows
```

```r
# Creating a new variable for cbind()
price_sqft <- housing_df$sale_price/housing_df$square_feet_total_living
price_sqft <- round(price_sqft, 2)
new_df <- cbind(housing_df,price_sqft)
# Using GROUP_BY to group by bedrooms and SUMMARIZE to find
# the mean price by bedroom and then ARRANGE by AvgPrice
housing_df %>% group_by(bedrooms) %>% summarize(AvgPrice=mean(sale_price)) %>%
  arrange(desc(AvgPrice))
```

```
## # A tibble: 12 x 2
##    bedrooms AvgPrice
##       <dbl>    <dbl>
##  1       11 1825000
##  2        7 1307282.
```

```
##  3        8 1122500
##  4        0  844059.
##  5        5  836974.
##  6        6  767494.
##  7        4  735910.
##  8        1  722814.
##  9        9  581500
## 10        3  564959.
## 11        2  544946.
## 12       10  450000
```

```
# Using KEEP()
even_sales <- keep(housing_df$sale_price, function(x) x %% 2 == 0)
# Using DISCARD()
odd_built_years <- discard(housing_df$year_built, function(x) x %% 2 == 0)
# Using MAP_INT
housing_df %>% map_int(NROW)
```

```
##               Sale Date              sale_price             sale_reason
##                   12865                   12865                   12865
##          sale_instrument            sale_warning                sitetype
##                   12865                   12865                   12865
##                addr_full                    zip5                 ctyname
##                   12865                   12865                   12865
##               postalctyn                     lon                     lat
##                   12865                   12865                   12865
##           building_grade square_feet_total_living                bedrooms
##                   12865                   12865                   12865
##          bath_full_count         bath_half_count         bath_3qtr_count
##                   12865                   12865                   12865
##               year_built          year_renovated          current_zoning
##                   12865                   12865                   12865
##                sq_ft_lot               prop_type             present_use
##                   12865                   12865                   12865
```

```
housing_df %>% map_int(NCOL)
```

```
##               Sale Date              sale_price             sale_reason
##                       1                       1                       1
##          sale_instrument            sale_warning                sitetype
##                       1                       1                       1
##                addr_full                    zip5                 ctyname
##                       1                       1                       1
##               postalctyn                     lon                     lat
##                       1                       1                       1
##           building_grade square_feet_total_living                bedrooms
##                       1                       1                       1
##          bath_full_count         bath_half_count         bath_3qtr_count
##                       1                       1                       1
##               year_built          year_renovated          current_zoning
##                       1                       1                       1
##                sq_ft_lot               prop_type             present_use
##                       1                       1                       1
```

```r
# SPLIT, CONCATENATE strings
addr_dir <- str_split(string = housing_df$addr_full, pattern = " ")
combined <- paste(addr_dir, sep = " ")
```