```
# Assignment: ASSIGNMENT 5.2
> # Name: Supraja, Rapuru
> # Date: 2021-07-10
> #Analysis of housing data
>
> ## Load the readxl package
> library(readxl)
> ## Load the dplyr package
> library(dplyr)
> ## Load the purrr package
> library(purrr)
>
>
> ## Set the working directory to the root of your DSC 520 directory
> setwd("/Users/Supraja/dsc520")
> ## Load the `data/acs-14-1yr-s0201.csv` to
> housing_df <- read_excel("data/week-6-housing.xlsx")
> 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
NE" "3303 178TH AVE NE" ...
 $ 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 ...
> head(housing_df)
# A tibble: 6 x 24
  `Sale Date`         `Sale Price` sale_reason sale_instrument sale_warning
  <dttm>                     <dbl>       <dbl>          <dbl> <chr>
1 2006-01-03 00:00:00       698000           1              3 NA
2 2006-01-03 00:00:00       649990           1              3 NA
3 2006-01-03 00:00:00       572500           1              3 NA
4 2006-01-03 00:00:00       420000           1              3 NA
5 2006-01-03 00:00:00       369900           1              3 15
6 2006-01-03 00:00:00       184667           1             15 18 51
# ... with 19 more variables: sitetype <chr>, addr_full <chr>, zip5 <dbl>,
#   ctyname <chr>, postalctyn <chr>, lon <dbl>, lat <dbl>,
#   building_grade <dbl>, square_feet_total_living <dbl>, bedrooms <dbl>,
```

```
#   bath_full_count <dbl>, bath_half_count <dbl>, bath_3qtr_count <dbl>,
#   year_built <dbl>, year_renovated <dbl>, current_zoning <chr>,
#   sq_ft_lot <dbl>, prop_type <chr>, present_use <dbl>
> #Rename the'Sale Date` and`Sale Price`
> colnames(housing_df)[1] <- "Sale_Date"
> colnames(housing_df)[2] <- "Sale_Price"
> 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
NE" "3303 178TH AVE NE" ...
 $ 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 …
> # a. Using the dplyr package, use the 6 different operations to analyze/transform
> # the data - GroupBy, Summarize, Mutate, Filter, Select, and Arrange – Remember
> # this isn’t just modifying data, you are learning about your data also – so play
> # around and start to understand your dataset in more detail
> #Getting mean sale price using group_by() and summarize() functions
> housing_df %>% group_by(zip5) %>% summarize("Avg_Sale_Price" = mean(Sale_Price))
# A tibble: 4 x 2
   zip5 Avg_Sale_Price
  <dbl>          <dbl>
1 98052         649375.
2 98053         672624.
3 98059         645000
4 98074         951544.
> #Getting mean sale price using group_by() and summarize() functions
> housing_df %>% group_by(zip5,ctyname) %>% summarize("Avg_Sale_Price" = mean(Sale_Price))
`summarise()` has grouped output by 'zip5'. You can override using the `.groups` argument.
# A tibble: 6 x 3
# Groups:   zip5 [4]
   zip5 ctyname   Avg_Sale_Price
  <dbl> <chr>              <dbl>
1 98052 REDMOND           644803.
2 98052 NA                691413.
3 98053 NA                672624.
4 98059 NA                645000
```

```
5 98074 SAMMAMISH       972480.
6 98074 NA              754143.
> #Getting mean sale price using group_by() and summarize() functions
> housing_df %>% group_by(bedrooms) %>% summarize("Avg_Sale_Price" = mean(Sale_Price))
# A tibble: 12 x 2
   bedrooms Avg_Sale_Price
      <dbl>          <dbl>
 1        0        844059.
 2        1        722814.
 3        2        544946.
 4        3        564959.
 5        4        735910.
 6        5        836974.
 7        6        767494.
 8        7       1307282.
 9        8       1122500
10        9        581500
11       10        450000
12       11       1825000
> #Getting mean sale price using group_by() and summarize() functions
> housing_df %>% group_by(year_built) %>% summarize("Avg_Sale_Price" = mean(Sale_Price))
# A tibble: 109 x 2
   year_built Avg_Sale_Price
        <dbl>          <dbl>
 1       1900        394500.
 2       1903        430000
 3       1905        620000
 4       1906        550000
 5       1909          1070
 6       1910        150000
 7       1912        619667.
 8       1913        457500
 9       1914        835000
10       1915        228150
# ... with 99 more rows
> #Calculate sales_price_per_sqft using mutate() function
> housing_df<-housing_df %>% mutate("sales_price_per_sqft"=square_feet_total_living/Sale_Price)
> str(housing_df)
tibble [12,865 x 25] (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
NE" "3303 178TH AVE NE" ...
 $ 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 ...
 $ sales_price_per_sqft    : num [1:12865] 0.00403 0.00443 0.00484 0.00386 0.00389 ...
> #Calculate sales_year using mutate() function
> housing_df<-housing_df %>% mutate("sale_year"=substr(Sale_Date,1,4))
> str(housing_df)
tibble [12,865 x 26] (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
NE" "3303 178TH AVE NE" ...
 $ 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 ...
 $ sales_price_per_sqft    : num [1:12865] 0.00403 0.00443 0.00484 0.00386 0.00389 ...
 $ sale_year             : chr [1:12865] "2006" "2006" "2006" "2006" ...
> #Filter all 4-bedroom houses using filter() function
> housing_df %>% filter(bedrooms==4)
# A tibble: 5,515 x 26
   Sale_Date           Sale_Price sale_reason sale_instrument sale_warning sitetype addr_full          zip5
ctyname postalctyn   lon   lat building_grade
   <dttm>                   <dbl>      <dbl>          <dbl> <chr>       <chr>    <chr>              <dbl> <chr>   <chr>
<dbl> <dbl>         <dbl>
 1 2006-01-03 00:00:00    698000          1              3 <NA>        R1       17021 NE 113TH CT  98052
REDMOND REDMOND    -122.  47.7             9
 2 2006-01-03 00:00:00    649990          1              3 <NA>        R1       11927 178TH PL NE  98052
REDMOND REDMOND    -122.  47.7             9
 3 2006-01-03 00:00:00    572500          1              3 <NA>        R1       13315 174TH AVE NE 98052 <NA>
REDMOND    -122.  47.7             8
 4 2006-01-03 00:00:00    184667          1             15 18 51       R1       8101 229TH DR NE   98053 <NA>
REDMOND    -122.  47.7             7
 5 2006-01-04 00:00:00    875000          1              3 <NA>        R1       21404 NE 67TH ST   98053 <NA>
REDMOND    -122.  47.7            10
```

```
 6 2006-01-04 00:00:00    660000        1          3 <NA>       R1      7525 238TH AVE NE  98053 <NA>
REDMOND    -122.  47.7          9
 7 2006-01-04 00:00:00    650000        1          3 <NA>       R1      17703 NE 26TH ST   98052
REDMOND REDMOND    -122.  47.6          8
 8 2006-01-04 00:00:00    470000        1          3 <NA>       R1      17905 NE 26TH ST   98052
REDMOND REDMOND    -122.  47.6          8
 9 2006-01-06 00:00:00    765000        1          3 <NA>       R1      8944 237TH PL NE   98053 <NA>
REDMOND    -122.  47.7          9
10 2006-01-06 00:00:00    589950        1          3 <NA>       R1      11922 173RD PL NE  98052
REDMOND REDMOND    -122.  47.7          8
# ... with 5,505 more rows, and 13 more variables: square_feet_total_living <dbl>, bedrooms <dbl>,
bath_full_count <dbl>, bath_half_count <dbl>,
#   bath_3qtr_count <dbl>, year_built <dbl>, year_renovated <dbl>, current_zoning <chr>, sq_ft_lot
<dbl>, prop_type <chr>, present_use <dbl>,
#   sales_price_per_sqft <dbl>, sale_year <chr>
> #Filter all houses whose sale price < 500000 using filter() function
> housing_df %>% filter(Sale_Price<500000)
# A tibble: 4,040 x 26
  Sale_Date          Sale_Price sale_reason sale_instrument sale_warning sitetype addr_full           zip5
ctyname postalctyn   lon   lat building_grade
  <dttm>                  <dbl>      <dbl>          <dbl> <chr>       <chr>    <chr>               <dbl> <chr>   <chr>
<dbl> <dbl>         <dbl>
 1 2006-01-03 00:00:00    420000        1          3 <NA>       R1      3303 178TH AVE NE    98052
REDMOND REDMOND    -122.  47.6          8
 2 2006-01-03 00:00:00    369900        1          3 15         R1      16126 NE 108TH CT    98052
REDMOND REDMOND    -122.  47.7          7
 3 2006-01-03 00:00:00    184667        1         15 18 51      R1      8101 229TH DR NE     98053 <NA>
REDMOND    -122.  47.7          7
 4 2006-01-04 00:00:00    470000        1          3 <NA>       R1      17905 NE 26TH ST     98052
REDMOND REDMOND    -122.  47.6          8
 5 2006-01-04 00:00:00    165000        1          3 <NA>       R1      2921 288TH AVE NE    98053
<NA>    REDMOND    -122.  47.6          9
 6 2006-01-09 00:00:00    372500        1          3 <NA>       R1      26920 NE 50TH ST     98053 <NA>
REDMOND    -122.  47.7          7
 7 2006-01-10 00:00:00    482000        1          3 <NA>       R1      9166 226TH PL NE     98053 <NA>
REDMOND    -122.  47.7          7
 8 2006-01-11 00:00:00    372500        1          3 <NA>       R2      8606 134TH CT NE     98052
REDMOND REDMOND    -122.  47.7          7
 9 2006-01-11 00:00:00    265000        1          3 <NA>       R1      25149 NE PATTERSON ~ 98053
<NA>    REDMOND    -122.  47.7         10
10 2006-01-12 00:00:00    470000        1          3 <NA>       R1      14876 NE 78TH WAY    98052
REDMOND REDMOND    -122.  47.7          8
# ... with 4,030 more rows, and 13 more variables: square_feet_total_living <dbl>, bedrooms <dbl>,
bath_full_count <dbl>, bath_half_count <dbl>,
#   bath_3qtr_count <dbl>, year_built <dbl>, year_renovated <dbl>, current_zoning <chr>, sq_ft_lot
<dbl>, prop_type <chr>, present_use <dbl>,
#   sales_price_per_sqft <dbl>, sale_year <chr>
> #Filter all houses which are sold in 2006 and sale price is less than 500000 using filter() function
> housing_df %>% filter(Sale_Price<500000& sale_year=='2006')
# A tibble: 524 x 26
  Sale_Date          Sale_Price sale_reason sale_instrument sale_warning sitetype addr_full           zip5
ctyname postalctyn   lon   lat building_grade
  <dttm>                  <dbl>      <dbl>          <dbl> <chr>       <chr>    <chr>               <dbl> <chr>   <chr>
<dbl> <dbl>         <dbl>
 1 2006-01-03 00:00:00    420000        1          3 <NA>       R1      3303 178TH AVE NE    98052
REDMOND REDMOND    -122.  47.6          8
```

```
 2 2006-01-03 00:00:00    369900        1          3 15        R1     16126 NE 108TH CT    98052
REDMOND REDMOND    -122.  47.7          7
 3 2006-01-03 00:00:00    184667        1         15 18 51     R1      8101 229TH DR NE     98053 <NA>
REDMOND    -122.  47.7          7
 4 2006-01-04 00:00:00    470000        1          3 <NA>      R1      17905 NE 26TH ST     98052
REDMOND REDMOND    -122.  47.6          8
 5 2006-01-04 00:00:00    165000        1          3 <NA>      R1      2921 288TH AVE NE    98053
<NA>    REDMOND    -122.  47.6          9
 6 2006-01-09 00:00:00    372500        1          3 <NA>      R1      26920 NE 50TH ST     98053 <NA>
REDMOND    -122.  47.7          7
 7 2006-01-10 00:00:00    482000        1          3 <NA>      R1      9166 226TH PL NE     98053 <NA>
REDMOND    -122.  47.7          7
 8 2006-01-11 00:00:00    372500        1          3 <NA>      R2      8606 134TH CT NE     98052
REDMOND REDMOND    -122.  47.7          7
 9 2006-01-11 00:00:00    265000        1          3 <NA>      R1      25149 NE PATTERSON ~ 98053
<NA>    REDMOND    -122.  47.7         10
10 2006-01-12 00:00:00    470000        1          3 <NA>      R1      14876 NE 78TH WAY    98052
REDMOND REDMOND    -122.  47.7          8
# ... with 514 more rows, and 13 more variables: square_feet_total_living <dbl>, bedrooms <dbl>,
bath_full_count <dbl>, bath_half_count <dbl>,
#   bath_3qtr_count <dbl>, year_built <dbl>, year_renovated <dbl>, current_zoning <chr>, sq_ft_lot
<dbl>, prop_type <chr>, present_use <dbl>,
#   sales_price_per_sqft <dbl>, sale_year <chr>
> #Select Sale_Date, sale_price and zip from the dataset using select() function
> housing_df %>% select(Sale_Date,Sale_Price,zip5)
# A tibble: 12,865 x 3
   Sale_Date          Sale_Price  zip5
   <dttm>                  <dbl> <dbl>
 1 2006-01-03 00:00:00    698000 98052
 2 2006-01-03 00:00:00    649990 98052
 3 2006-01-03 00:00:00    572500 98052
 4 2006-01-03 00:00:00    420000 98052
 5 2006-01-03 00:00:00    369900 98052
 6 2006-01-03 00:00:00    184667 98053
 7 2006-01-04 00:00:00   1050000 98053
 8 2006-01-04 00:00:00    875000 98053
 9 2006-01-04 00:00:00    660000 98053
10 2006-01-04 00:00:00    650000 98052
# ... with 12,855 more rows
> #Select Sale_Date, sale_price and zip from the dataset for 11-bedroom house using filter() and
select() function
> housing_df %>% filter(bedrooms==11)%>% select(Sale_Date,Sale_Price,zip5)
# A tibble: 1 x 3
  Sale_Date          Sale_Price  zip5
  <dttm>                  <dbl> <dbl>
1 2007-12-11 00:00:00   1825000 98052
> #Arrange the dataset based on sales price from high to low
> housing_df %>% arrange(desc(Sale_Price))
# A tibble: 12,865 x 26
   Sale_Date          Sale_Price sale_reason sale_instrument sale_warning sitetype addr_full          zip5
ctyname postalctyn   lon   lat building_grade
   <dttm>                  <dbl>      <dbl>          <dbl> <chr>       <chr>    <chr>              <dbl> <chr>   <chr>
<dbl> <dbl>         <dbl>
 1 2010-03-02 00:00:00   4400000          1              3 35 45       R1       12025 154TH PL NE  98052 <NA>
REDMOND    -122.  47.7          11
 2 2010-03-02 00:00:00   4400000          1              3 35 45       R1       12053 154TH PL NE  98052 <NA>
REDMOND    -122.  47.7           6
```

```
 3 2011-11-17 00:00:00   4380542        1         22 11 45      R1      17137 NE 120TH ST  98052
REDMOND REDMOND   -122.  47.7          8
 4 2011-11-17 00:00:00   4380542        1         22 11 45      R1      11818 171ST PL NE  98052
REDMOND REDMOND   -122.  47.7          8
 5 2011-11-17 00:00:00   4380542        1         22 11 45      R1      17011 NE 118TH WAY 98052
REDMOND REDMOND   -122.  47.7          8
 6 2011-11-17 00:00:00   4380542        1         22 11 45      R1      16943 NE 118TH WAY 98052
REDMOND REDMOND   -122.  47.7          8
 7 2011-11-17 00:00:00   4380542        1         22 11 45      R1      16944 NE 118TH WAY 98052
REDMOND REDMOND   -122.  47.7          8
 8 2011-11-17 00:00:00   4380542        1         22 11 45      R1      16909 NE 120TH ST  98052
REDMOND REDMOND   -122.  47.7          8
 9 2011-11-17 00:00:00   4380542        1         22 11 45      R1      17128 NE 120TH ST  98052
REDMOND REDMOND   -122.  47.7          8
10 2011-11-17 00:00:00   4380542        1         22 11 45      R1      17136 NE 120TH ST  98052
REDMOND REDMOND   -122.  47.7          8
# ... with 12,855 more rows, and 13 more variables: square_feet_total_living <dbl>, bedrooms <dbl>,
bath_full_count <dbl>, bath_half_count <dbl>,
#   bath_3qtr_count <dbl>, year_built <dbl>, year_renovated <dbl>, current_zoning <chr>, sq_ft_lot
<dbl>, prop_type <chr>, present_use <dbl>,
#   sales_price_per_sqft <dbl>, sale_year <chr>
> # b.Using the purrr package – perform 2 functions on your dataset.
> # You could use zip_n, keep, discard, compact, etc.
> #Using keep function list all the sales prices which are greater than 2000000
> sales_price_gt_2m <-purrr::keep(housing_df$Sale_Price, ~ .x>2000000)
> class(sales_price_gt_2m)
[1] "numeric"
> str(sales_price_gt_2m)
 num [1:206] 2500000 2169000 2569000 2583000 3000000 ...
> # b.Using the purrr package – perform 2 functions on your dataset.
> # You could use zip_n, keep, discard, compact, etc.
> #Using keep function list all the sales prices which are greater than 2000000
> sales_price_gt_2m <-purrr::keep(housing_df$Sale_Price, ~ .x>2000000)
> class(sales_price_gt_2m)
[1] "numeric"
> str(sales_price_gt_2m)
 num [1:206] 2500000 2169000 2569000 2583000 3000000 ...
> #Perform map function on the list to generate a list with sales price increased by 5%
> sales_price_gt_2m %>% map(function(x) x*.05)
[[1]]
[1] 125000

[[2]]
[1] 108450

[[3]]
[1] 128450

[[4]]
[1] 129150

[[5]]
[1] 150000

[[6]]
[1] 111750
```

[[7]]
[1] 149400

[[8]]
[1] 124650

[[9]]
[1] 131250

[[10]]
[1] 131250

[[11]]
[1] 131250

[[12]]
[1] 131250

[[13]]
[1] 131250

[[14]]
[1] 131250

[[15]]
[1] 131250

[[16]]
[1] 129500

[[17]]
[1] 129500

[[18]]
[1] 129500

[[19]]
[1] 129500

[[20]]
[1] 129500

[[21]]
[1] 129500

[[22]]
[1] 129500

[[23]]
[1] 115000

[[24]]
[1] 115000

[[25]]
[1] 115000

[[26]]
[1] 129900

[[27]]
[1] 199750

[[28]]
[1] 104078.6

[[29]]
[1] 127450

[[30]]
[1] 104000

[[31]]
[1] 109450

[[32]]
[1] 158750

[[33]]
[1] 158750

[[34]]
[1] 158750

[[35]]
[1] 158750

[[36]]
[1] 158750

[[37]]
[1] 158750

[[38]]
[1] 158750

[[39]]
[1] 158750

[[40]]
[1] 158750

[[41]]
[1] 158750

[[42]]
[1] 158750

[[43]]
[1] 158750

[[44]]
[1] 158750

[[45]]
[1] 158750

[[46]]
[1] 158750

[[47]]
[1] 158750

[[48]]
[1] 158750

[[49]]
[1] 158750

[[50]]
[1] 158750

[[51]]
[1] 158750

[[52]]
[1] 158750

[[53]]
[1] 158750

[[54]]
[1] 158750

[[55]]
[1] 158750

[[56]]
[1] 158750

[[57]]
[1] 158750

[[58]]
[1] 158750

[[59]]
[1] 158750

[[60]]
[1] 158750

[[61]]
[1] 158750

[[62]]
[1] 158750

[[63]]
[1] 158750

[[64]]
[1] 158750

[[65]]
[1] 158750

[[66]]
[1] 158750

[[67]]
[1] 157500

[[68]]
[1] 157500

[[69]]
[1] 157500

[[70]]
[1] 157500

[[71]]
[1] 157500

[[72]]
[1] 157500

[[73]]
[1] 157500

[[74]]
[1] 157500

[[75]]
[1] 157500

[[76]]
[1] 157500

[[77]]
[1] 157500

[[78]]
[1] 157500

[[79]]
[1] 157500

[[80]]
[1] 157500

[[81]]
[1] 157500

[[82]]
[1] 157500

[[83]]
[1] 157500

[[84]]
[1] 157500

[[85]]
[1] 157500

[[86]]
[1] 157500

[[87]]
[1] 157500

[[88]]
[1] 157500

[[89]]
[1] 157500

[[90]]
[1] 157500

[[91]]
[1] 157500

[[92]]
[1] 157500

[[93]]
[1] 157500

[[94]]
[1] 157500

[[95]]
[1] 157500

[[96]]
[1] 157500

[[97]]
[1] 157500

[[98]]
[1] 157500

[[99]]
[1] 157500

[[100]]
[1] 157500

[[101]]
[1] 101650

[[102]]
[1] 220000

[[103]]
[1] 220000

[[104]]
[1] 115000

[[105]]
[1] 115000

[[106]]
[1] 115000

[[107]]
[1] 144250

[[108]]
[1] 144250

[[109]]
[1] 144250

[[110]]
[1] 144250

[[111]]
[1] 144250

[[112]]
[1] 144250

[[113]]
[1] 144250

[[114]]
[1] 144250

[[115]]
[1] 144250

[[116]]
[1] 144250

[[117]]
[1] 219027.1

[[118]]
[1] 219027.1

[[119]]
[1] 219027.1

[[120]]
[1] 219027.1

[[121]]
[1] 219027.1

[[122]]
[1] 219027.1

[[123]]
[1] 219027.1

[[124]]
[1] 219027.1

[[125]]
[1] 219027.1

[[126]]
[1] 219027.1

[[127]]
[1] 219027.1

[[128]]
[1] 219027.1

[[129]]
[1] 219027.1

[[130]]
[1] 219027.1

[[131]]
[1] 207010.2

[[132]]
[1] 207010.2

[[133]]
[1] 207010.2

[[134]]
[1] 207010.2

[[135]]
[1] 207010.2

[[136]]
[1] 207010.2

[[137]]
[1] 207010.2

[[138]]
[1] 207010.2

[[139]]
[1] 207010.2

[[140]]
[1] 207010.2

[[141]]
[1] 207010.2

[[142]]
[1] 207010.2

[[143]]
[1] 207010.2

[[144]]
[1] 207010.2

[[145]]
[1] 207010.2

[[146]]
[1] 125000

[[147]]
[1] 115000

[[148]]
[1] 115000

[[149]]
[1] 115000

[[150]]
[1] 115000

[[151]]
[1] 115000

[[152]]
[1] 115000

[[153]]
[1] 115000

[[154]]
[1] 115000

[[155]]
[1] 115000

[[156]]
[1] 115000

[[157]]
[1] 115000

[[158]]
[1] 160000

[[159]]
[1] 125000

[[160]]
[1] 125000

[[161]]
[1] 173100

[[162]]
[1] 173100

[[163]]
[1] 173100

[[164]]
[1] 173100

[[165]]
[1] 173100

[[166]]
[1] 173100

[[167]]
[1] 173100

[[168]]
[1] 173100

[[169]]
[1] 173100

[[170]]
[1] 173100

[[171]]
[1] 173100

[[172]]
[1] 173100

[[173]]
[1] 173100

[[174]]
[1] 173100

[[175]]
[1] 173100

[[176]]
[1] 173100

[[177]]
[1] 173100

[[178]]
[1] 173100

[[179]]
[1] 115000

[[180]]
[1] 150000

[[181]]
[1] 124557.5

[[182]]
[1] 137500

[[183]]
[1] 167000

[[184]]
[1] 108010

[[185]]
[1] 108010

[[186]]
[1] 114000

[[187]]
[1] 110000

[[188]]
[1] 107000

[[189]]
[1] 110000

[[190]]
[1] 115000

[[191]]
[1] 101250

[[192]]
[1] 107500

[[193]]
[1] 187500

[[194]]
[1] 142500

[[195]]
[1] 108250

[[196]]
[1] 108250

```
[[197]]
[1] 158750

[[198]]
[1] 215550

[[199]]
[1] 103850

[[200]]
[1] 117500

[[201]]
[1] 110000

[[202]]
[1] 135000

[[203]]
[1] 197500

[[204]]
[1] 192500

[[205]]
[1] 149400

[[206]]
[1] 102500
> #Using discard function list all the sale year which are greater than 2000
> sale_year_gt_2000<-purrr::discard(housing_df$sale_year, ~ .x<2000)
> class(sale_year_gt_2000)
[1] "character"
> str(sale_year_gt_2000)
 chr [1:12865] "2006" "2006" "2006" "2006" "2006" "2006" "2006" "2006" "2006" "2006" "2006"
"2006" "2006" "2006" "2006" "2006" "2006" "2006" "2006" ...
> unique(sale_year_gt_2000)
 [1] "2006" "2007" "2008" "2009" "2010" "2011" "2012" "2013" "2014" "2015" "2016"
> # c.Use the cbind and rbind function on your dataset
> #using cbind function add city_indicator
> housing_df <-cbind(housing_df,city_indicator=!is.na(housing_df$ctyname))
> str(housing_df)
'data.frame':   12865 obs. of  27 variables:
 $ Sale_Date            : POSIXct, format: "2006-01-03" "2006-01-03" "2006-01-03" "2006-01-03" ...
 $ Sale_Price           : num  698000 649990 572500 420000 369900 ...
 $ sale_reason           : num  1 1 1 1 1 1 1 1 1 1 ...
 $ sale_instrument       : num  3 3 3 3 3 15 3 3 3 3 ...
 $ sale_warning          : chr  NA NA NA NA ...
 $ sitetype             : chr  "R1" "R1" "R1" "R1" ...
 $ addr_full            : chr  "17021 NE 113TH CT" "11927 178TH PL NE" "13315 174TH AVE NE" "3303
178TH AVE NE" ...
 $ zip5                 : num  98052 98052 98052 98052 98052 ...
 $ ctyname              : chr  "REDMOND" "REDMOND" NA "REDMOND" ...
 $ postalctyn           : chr  "REDMOND" "REDMOND" "REDMOND" "REDMOND" ...
 $ lon                  : num  -122 -122 -122 -122 -122 ...
 $ lat                  : num  47.7 47.7 47.7 47.6 47.7 ...
 $ building_grade        : num  9 9 8 8 7 7 10 10 9 8 ...
```

```
 $ square_feet_total_living: num  2810 2880 2770 1620 1440 4160 3960 3720 4160 2760 ...
 $ bedrooms             : num  4 4 4 3 3 4 5 4 4 4 ...
 $ bath_full_count      : num  2 2 1 1 1 2 3 2 2 1 ...
 $ bath_half_count      : num  1 0 1 0 0 1 0 1 1 0 ...
 $ bath_3qtr_count      : num  0 1 1 1 1 1 1 0 1 1 ...
 $ year_built           : num  2003 2006 1987 1968 1980 ...
 $ year_renovated       : num  0 0 0 0 0 0 0 0 0 0 ...
 $ current_zoning       : chr  "R4" "R4" "R6" "R4" ...
 $ sq_ft_lot            : num  6635 5570 8444 9600 7526 ...
 $ prop_type            : chr  "R" "R" "R" "R" ...
 $ present_use          : num  2 2 2 2 2 2 2 2 2 2 ...
 $ sales_price_per_sqft    : num  0.00403 0.00443 0.00484 0.00386 0.00389 ...
 $ sale_year            : chr  "2006" "2006" "2006" "2006" ...
 $ city_indicator       : logi  TRUE TRUE FALSE TRUE TRUE FALSE ...
> #Using rbind function to combine 2 dataframes
> hs_sale_yr_bfr_2010<-housing_df %>%filter(sale_year<2010)
> head(hs_sale_yr_bfr_2010)
  Sale_Date Sale_Price sale_reason sale_instrument sale_warning sitetype        addr_full  zip5
ctyname postalctyn      lon     lat building_grade
1 2006-01-03     698000         1           3       <NA>      R1  17021 NE 113TH CT 98052 REDMOND
REDMOND -122.1124 47.70139           9
2 2006-01-03     649990         1           3       <NA>      R1  11927 178TH PL NE 98052 REDMOND
REDMOND -122.1022 47.70731           9
3 2006-01-03     572500         1           3       <NA>      R1 13315 174TH AVE NE 98052    <NA>
REDMOND -122.1085 47.71986           8
4 2006-01-03     420000         1           3       <NA>      R1  3303 178TH AVE NE 98052 REDMOND
REDMOND -122.1037 47.63914           8
5 2006-01-03     369900         1           3         15      R1  16126 NE 108TH CT 98052 REDMOND
REDMOND -122.1242 47.69748           7
6 2006-01-03     184667         1          15       18 51      R1   8101 229TH DR NE 98053    <NA>
REDMOND -122.0341 47.67545           7
  square_feet_total_living bedrooms bath_full_count bath_half_count bath_3qtr_count year_built
year_renovated current_zoning sq_ft_lot prop_type
1                     2810        4               2               1               0       2003
0               R4       6635         R
2                     2880        4               2               0               1       2006
0               R4       5570         R
3                     2770        4               1               1               1       1987
0               R6       8444         R
4                     1620        3               1               0               1       1968
0               R4       9600         R
5                     1440        3               1               0               1       1980
0               R6       7526         R
6                     4160        4               2               1               1       2005
0            URPSO       7280         R
  present_use sales_price_per_sqft sale_year city_indicator
1           2          0.004025788      2006           TRUE
2           2          0.004430837      2006           TRUE
3           2          0.004838428      2006          FALSE
4           2          0.003857143      2006           TRUE
5           2          0.003892944      2006           TRUE
6           2          0.022527035      2006          FALSE
> hs_sale_yr_aftr_2010<-housing_df %>%filter(sale_year>=2010)
> head(hs_sale_yr_aftr_2010)
  Sale_Date Sale_Price sale_reason sale_instrument sale_warning sitetype        addr_full  zip5
ctyname postalctyn      lon     lat building_grade
1 2010-01-04     750000         1           3         26      R1   19736 NE 61ST PL 98053    <NA>
REDMOND -122.0757 47.66093          11
2 2010-01-04     505000         1          22         46      R1  7220 218TH AVE NE 98053    <NA>
REDMOND -122.0481 47.66940           8
3 2010-01-04     155000         1           3         22      R1   9727 163RD PL NE 98052 REDMOND
REDMOND -122.1231 47.68738           8
```

```
4 2010-01-05     375000         1          3       <NA>      R1 23670 NE 135TH WAY 98053    <NA>
REDMOND -122.0223 47.71995          8
5 2010-01-06     540000         1          3       <NA>      R1  8220 208TH AVE NE 98053    <NA>
REDMOND -122.0608 47.67716          9
6 2010-01-06     540000        18         22       <NA>      R1   9879 187TH CT NE 98052 REDMOND
REDMOND -122.0909 47.68706          9
  square_feet_total_living bedrooms bath_full_count bath_half_count bath_3qtr_count year_built
year_renovated current_zoning sq_ft_lot prop_type
1                4250        4               2               1               1          2007
0          RA5    223027         R
2                3620        4               2               1               1          1987
0          RA5     37163         R
3                2250        4               1               0               2          1974
0           R5      8400         R
4                1340        2               2               0               0          2006
0         URPSO      4834         R
5                3060        5               1               0               2          1962
0          RA5    102847         R
6                2870        4               2               1               0          2006
0           R4      5409         R
  present_use sales_price_per_sqft sale_year city_indicator
1           2         0.005666667      2010         FALSE
2           2         0.007168317      2010         FALSE
3           2         0.014516129      2010          TRUE
4          29          0.003573333      2010         FALSE
5           2         0.005666667      2010         FALSE
6           2         0.005314815      2010          TRUE
> new_housing_df<-rbind(hs_sale_yr_bfr_2010,hs_sale_yr_aftr_2010)
> head(new_housing_df)
   Sale_Date Sale_Price sale_reason sale_instrument sale_warning sitetype         addr_full  zip5
ctyname postalctyn       lon      lat building_grade
1 2006-01-03     698000         1          3       <NA>      R1  17021 NE 113TH CT 98052 REDMOND
REDMOND -122.1124 47.70139          9
2 2006-01-03     649990         1          3       <NA>      R1  11927 178TH PL NE 98052 REDMOND
REDMOND -122.1022 47.70731          9
3 2006-01-03     572500         1          3       <NA>      R1 13315 174TH AVE NE 98052    <NA>
REDMOND -122.1085 47.71986          8
4 2006-01-03     420000         1          3       <NA>      R1  3303 178TH AVE NE 98052 REDMOND
REDMOND -122.1037 47.63914          8
5 2006-01-03     369900         1          3         15      R1  16126 NE 108TH CT 98052 REDMOND
REDMOND -122.1242 47.69748          7
6 2006-01-03     184667         1         15      18 51      R1   8101 229TH DR NE 98053    <NA>
REDMOND -122.0341 47.67545          7
  square_feet_total_living bedrooms bath_full_count bath_half_count bath_3qtr_count year_built
year_renovated current_zoning sq_ft_lot prop_type
1                2810        4               2               1               0          2003
0           R4      6635         R
2                2880        4               2               0               1          2006
0           R4      5570         R
3                2770        4               1               1               1          1987
0           R6      8444         R
4                1620        3               1               0               1          1968
0           R4      9600         R
5                1440        3               1               0               1          1980
0           R6      7526         R
6                4160        4               2               1               1          2005
0         URPSO      7280         R
  present_use sales_price_per_sqft sale_year city_indicator
1           2         0.004025788      2006          TRUE
2           2         0.004430837      2006          TRUE
3           2         0.004838428      2006         FALSE
4           2         0.003857143      2006          TRUE
5           2         0.003892944      2006          TRUE
6           2         0.022527035      2006         FALSE
> identical(new_housing_df,housing_df)
[1] TRUE
> # d.Split a string, then concatenate the results back together
> library(stringr)
```

```
> #split the Sale_Date columns
> sales_date_list<-str_split(string=housing_df$Sale_Date,pattern = '-')
> head(sales_date_list)
[[1]]
[1] "2006" "01"   "03"

[[2]]
[1] "2006" "01"   "03"

[[3]]
[1] "2006" "01"   "03"

[[4]]
[1] "2006" "01"   "03"

[[5]]
[1] "2006" "01"   "03"

[[6]]
[1] "2006" "01"   "03"
> #Create dataframe from the list
> sales_date_matrix=data.frame(Reduce(rbind,sales_date_list))
> head(sales_date_matrix)
      X1 X2 X3
init 2006 01 03
X    2006 01 03
X.1  2006 01 03
X.2  2006 01 03
X.3  2006 01 03
X.4  2006 01 03
> #assign names to the new columns
> names(sales_date_matrix)<- c('sale_year','sale_month','sale_date')
> head(sales_date_matrix)
    sale_year sale_month sale_date
init      2006        01       03
X         2006        01       03
X.1       2006        01       03
X.2       2006        01       03
X.3       2006        01       03
X.4       2006        01       03
> #combine the housing dataframe with new dataframe
> housing_df<-cbind(housing_df,sales_date_matrix)
> head(housing_df)
     Sale_Date Sale_Price sale_reason sale_instrument sale_warning sitetype         addr_full  zip5
ctyname postalctyn      lon      lat building_grade
init 2006-01-03     698000          1              3        <NA>       R1  17021 NE 113TH CT 98052 REDMOND
REDMOND -122.1124 47.70139             9
X    2006-01-03     649990          1              3        <NA>       R1  11927 178TH PL NE 98052 REDMOND
REDMOND -122.1022 47.70731             9
X.1  2006-01-03     572500          1              3        <NA>       R1 13315 174TH AVE NE 98052    <NA>
REDMOND -122.1085 47.71986             8
X.2  2006-01-03     420000          1              3        <NA>       R1  3303 178TH AVE NE 98052 REDMOND
REDMOND -122.1037 47.63914             8
X.3  2006-01-03     369900          1              3          15       R1  16126 NE 108TH CT 98052 REDMOND
REDMOND -122.1242 47.69748             7
X.4  2006-01-03     184667          1             15       18 51       R1   8101 229TH DR NE 98053    <NA>
REDMOND -122.0341 47.67545             7
```

	square_feet_total_living	bedrooms	bath_full_count	bath_half_count	bath_3qtr_count	year_built	year_renovated	current_zoning	sq_ft_lot	prop_type
init	2810	4	2	1	0	2003	0	R4	6635	R
X	2880	4	2	0	1	2006	0	R4	5570	R
X.1	2770	4	1	1	1	1987	0	R6	8444	R
X.2	1620	3	1	0	1	1968	0	R4	9600	R
X.3	1440	3	1	0	1	1980	0	R6	7526	R
X.4	4160	4	2	1	1	2005	0	URPSO	7280	R

	present_use	sales_price_per_sqft	sale_year	city_indicator	sale_year	sale_month	sale_date
init	2	0.004025788	2006	TRUE	2006	01	03
X	2	0.004430837	2006	TRUE	2006	01	03
X.1	2	0.004838428	2006	FALSE	2006	01	03
X.2	2	0.003857143	2006	TRUE	2006	01	03
X.3	2	0.003892944	2006	TRUE	2006	01	03
X.4	2	0.022527035	2006	FALSE	2006	01	03