# Assignment: ASSIGNMENT 4.2

# Name: Anjale, Jiteshwar

# Date: 2021-04-08

#Analysis of housing data

## Load the readxl package

library(readxl)

## Load the plyr package

library(plyr)

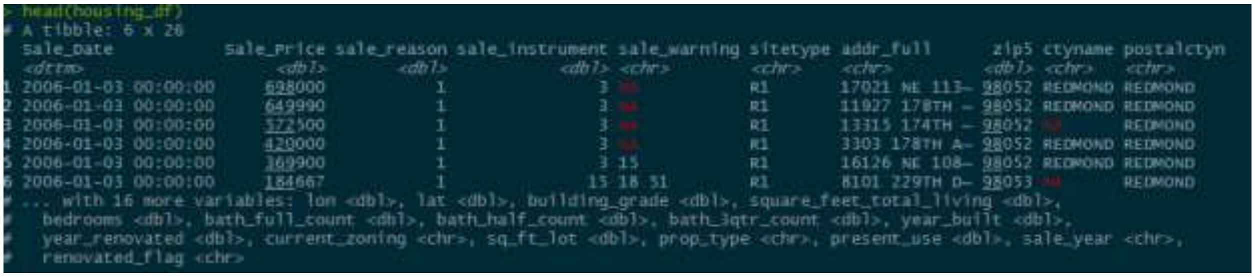
## Set the working directory to the root of your DSC 520 directory

setwd('C:/Users/anjal/OneDrive/Desktop/MS/DSC520/dsc520')

## Load the `data/acs-14-1yr-s0201.csv` to

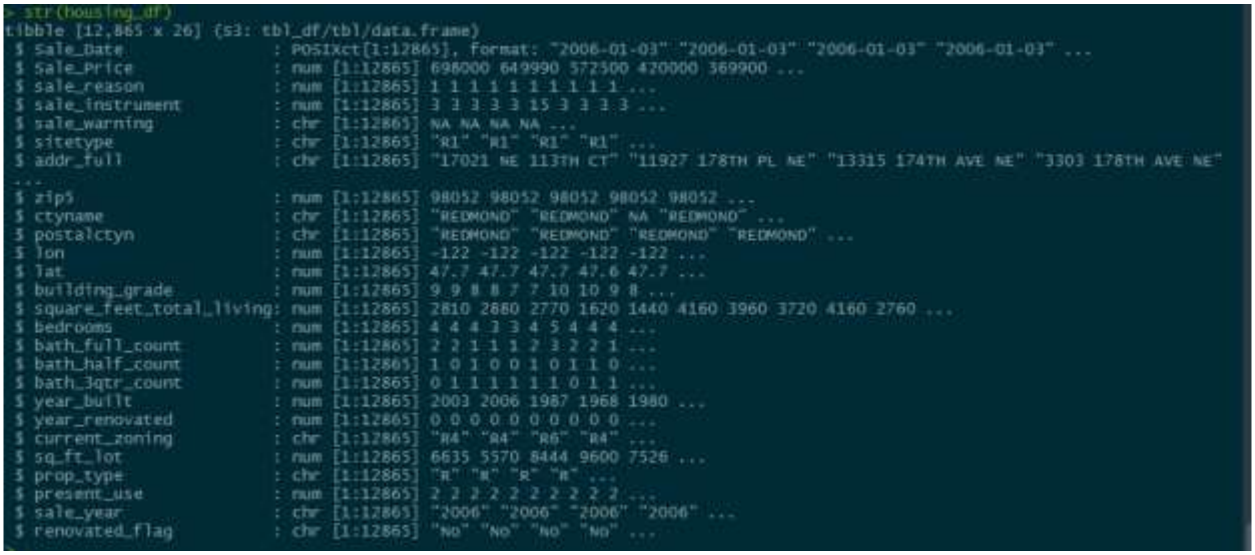
housing_df <- read_excel("data/week-6-housing.xlsx")

head(housing_df)



```
> head(housing_df)
# A tibble: 6 x 26
  Sale_Date           Sale_Price sale_reason sale_instrument sale_warning sitetype addr_full      zip5 ctyname  postalctyn
  <dttm>                   <dbl>       <dbl>           <dbl> <chr>        <chr>    <chr>         <dbl> <chr>    <chr>
1 2006-01-03 00:00:00     698000           1               3 NA           R1       17021 NE 113… 98052 REDMOND  REDMOND
2 2006-01-03 00:00:00     649990           1               3 NA           R1       11927 178TH … 98052 REDMOND  REDMOND
3 2006-01-03 00:00:00     572500           1               3 NA           R1       13315 174TH … 98052 NA       REDMOND
4 2006-01-03 00:00:00     420000           1               3 NA           R1       3303 178TH A… 98052 REDMOND  REDMOND
5 2006-01-03 00:00:00     369900           1               3 15           R1       16126 NE 108… 98052 REDMOND  REDMOND
6 2006-01-03 00:00:00     184667           1              15 18 51        R1       8101 229TH D… 98053 NA       REDMOND
# … with 16 more variables: 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>, sale_year <chr>,
#   renovated_flag <chr>
```

str(housing_df)

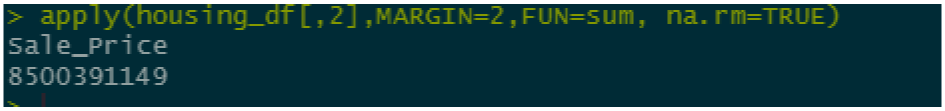


```
> str(housing_df)
tibble [12,865 x 26] (S3: tbl_df/tbl/data.frame)
 $ Sale_Date               : POSIXct[1:12865], format: "2006-01-03" "2006-01-03" "2006-01-03" "2006-01-03" ...
 $ Sale_Price              : num [1:12865] 698000 649990 572500 420000 369900 ...
 $ sale_reason             : num [1:12865] 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 ...
 $ sale_year               : chr [1:12865] "2006" "2006" "2006" "2006" ...
 $ renovated_flag          : chr [1:12865] "No" "No" "No" "No" ...
```

#a.Use the apply function on a variable in your dataset

#get sum of the sale price using apply function

apply(housing_df[,2],MARGIN=2,FUN=sum, na.rm=TRUE)



```
> apply(housing_df[,2],MARGIN=2,FUN=sum, na.rm=TRUE)
Sale_Price
8500391149
```

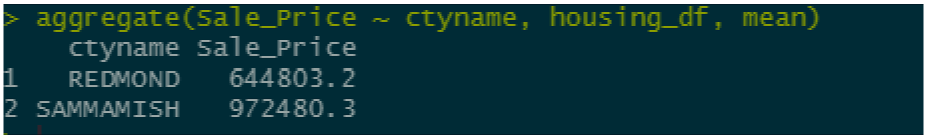
# b.Use the aggregate function on a variable in your dataset

#get mean sales price by cityname using aggregate function

colnames(housing_df)[1] <- "Sale_Date"

colnames(housing_df)[2] <- "Sale_Price"

aggregate(Sale_Price ~ ctyname, housing_df, mean)



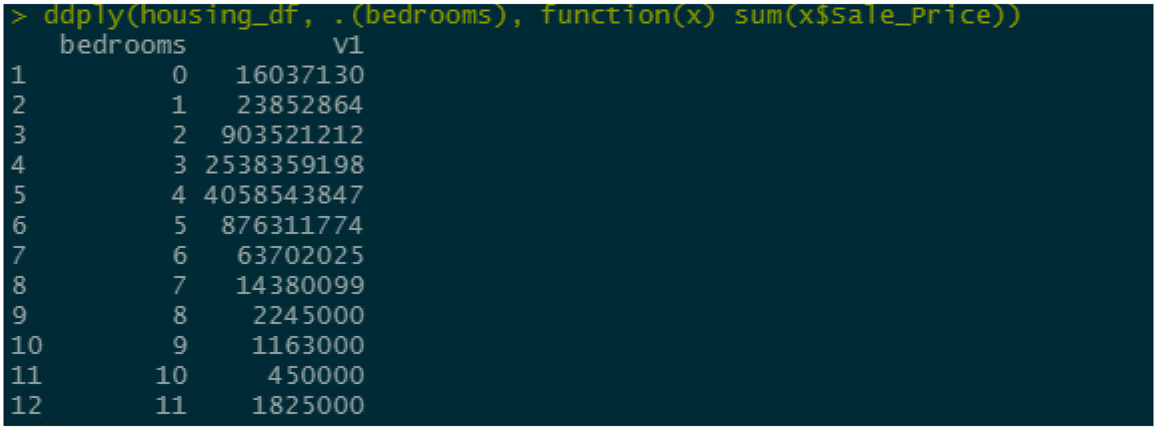
```
> aggregate(Sale_Price ~ ctyname, housing_df, mean)
    ctyname Sale_Price
1   REDMOND   644803.2
2 SAMMAMISH   972480.3
```

# c.Use the plyr function on a variable in your

# dataset – more specifically, I want to see you split some data,

# perform a modification to the data, and then bring it back together

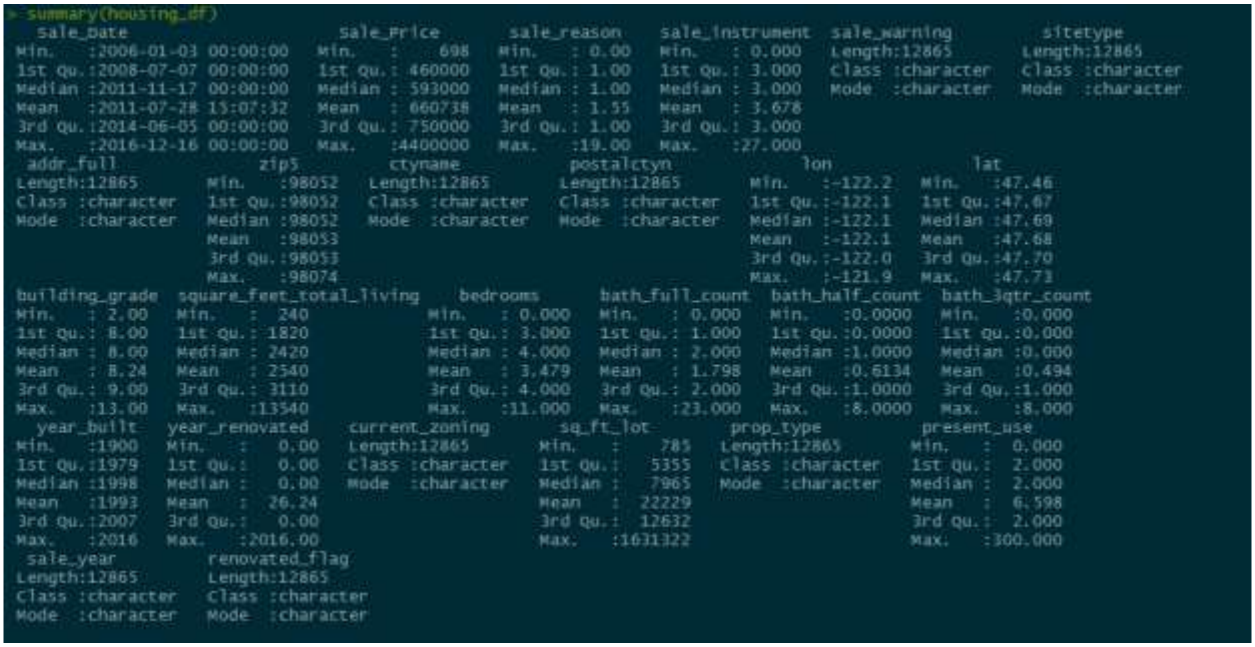
ddply(housing_df, .(bedrooms), function(x) sum(x$Sale_Price))



```
> ddply(housing_df, .(bedrooms), function(x) sum(x$Sale_Price))
   bedrooms         V1
1         0   16037130
2         1   23852864
3         2  903521212
4         3 2538359198
5         4 4058543847
6         5  876311774
7         6   63702025
8         7   14380099
9         8    2245000
10        9    1163000
11       10     450000
12       11    1825000
```

# d.Check distributions of the data

summary(housing_df)



```
> summary(housing_df)
   Sale_Date                      Sale_Price       sale_reason      sale_instrument  sale_warning          sitetype        
 Min.   :2006-01-03 00:00:00   Min.   :    698   Min.   : 0.00   Min.   : 0.000   Length:12865       Length:12865      
 1st Qu.:2008-07-07 00:00:00   1st Qu.: 460000   1st Qu.: 1.00   1st Qu.: 3.000   Class :character   Class :character  
 Median :2011-11-17 00:00:00   Median : 593000   Median : 1.00   Median : 3.000   Mode  :character   Mode  :character  
 Mean   :2011-07-28 15:07:32   Mean   : 660738   Mean   : 1.55   Mean   : 3.678                                        
 3rd Qu.:2014-06-05 00:00:00   3rd Qu.: 750000   3rd Qu.: 1.00   3rd Qu.: 3.000                                        
 Max.   :2016-12-16 00:00:00   Max.   :4400000   Max.   :19.00   Max.   :27.000                                        
  addr_full              zip5          ctyname           postalctyn             lon              lat       
 Length:12865       Min.   :98052   Length:12865       Length:12865       Min.   :-122.2   Min.   :47.46  
 Class :character   1st Qu.:98052   Class :character   Class :character   1st Qu.:-122.1   1st Qu.:47.67  
 Mode  :character   Median :98052   Mode  :character   Mode  :character   Median :-122.1   Median :47.69  
                    Mean   :98053                                         Mean   :-122.1   Mean   :47.68  
                    3rd Qu.:98053                                         3rd Qu.:-122.0   3rd Qu.:47.70  
                    Max.   :98074                                         Max.   :-121.9   Max.   :47.73  
 building_grade  square_feet_total_living    bedrooms      bath_full_count  bath_half_count  bath_3qtr_count
 Min.   : 2.00   Min.   :  240            Min.   : 0.000   Min.   : 0.000   Min.   :0.0000   Min.   :0.000  
 1st Qu.: 8.00   1st Qu.: 1820            1st Qu.: 3.000   1st Qu.: 1.000   1st Qu.:0.0000   1st Qu.:0.000  
 Median : 8.00   Median : 2420            Median : 4.000   Median : 2.000   Median :1.0000   Median :0.000  
 Mean   : 8.24   Mean   : 2540            Mean   : 3.479   Mean   : 1.798   Mean   :0.6134   Mean   :0.494  
 3rd Qu.: 9.00   3rd Qu.: 3110            3rd Qu.: 4.000   3rd Qu.: 2.000   3rd Qu.:1.0000   3rd Qu.:1.000  
 Max.   :13.00   Max.   :13540            Max.   :11.000   Max.   :23.000   Max.   :8.0000   Max.   :8.000  
   year_built   year_renovated    current_zoning        sq_ft_lot         prop_type          present_use     
 Min.   :1900   Min.   :   0.00   Length:12865       Min.   :    785   Length:12865       Min.   :  0.000  
 1st Qu.:1979   1st Qu.:   0.00   Class :character   1st Qu.:   5355   Class :character   1st Qu.:  2.000  
 Median :1998   Median :   0.00   Mode  :character   Median :   7965   Mode  :character   Median :  2.000  
 Mean   :1993   Mean   :  26.24                      Mean   :  22229                      Mean   :  6.598  
 3rd Qu.:2007   3rd Qu.:   0.00                      3rd Qu.:  12632                      3rd Qu.:  2.000  
 Max.   :2016   Max.   :2016.00                      Max.   :1631322                      Max.   :300.000  
  sale_year         renovated_flag    
 Length:12865       Length:12865      
 Class :character   Class :character  
 Mode  :character   Mode  :character  
```

#Sale_Price varies between 698 to 4400000. The mean Sale_Price is 660738.

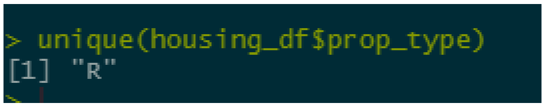
#Bedrooms varies between 0 to 11. There are so many variants available.

#year_built varies from 1900 to 2016. Some houses are very old available for sale.

#sq_ft_lot varies between 785 to 1631322.

#sale_Date varies between 2006-01-03 to 2016-12-16.

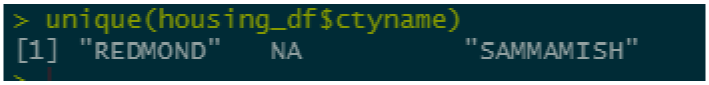
unique(housing_df$prop_type)



```
> unique(housing_df$prop_type)
[1] "R"
>
```

#All the houses are of type "R" i.e residential.

unique(housing_df$ctyname)



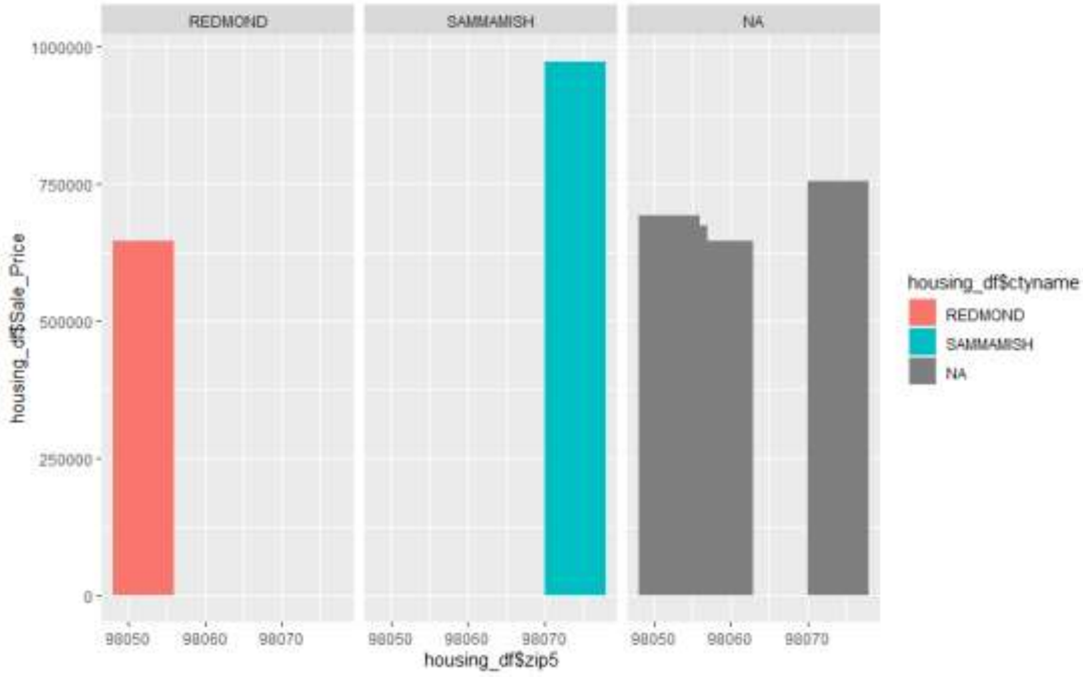
```
> unique(housing_df$ctyname)
[1] "REDMOND"    NA           "SAMMAMISH"
>
```

#All the houses are located in "REDMOND" and"SAMMAMISH"

library(ggplot2)

bar <- ggplot(housing_df, aes(housing_df$zip5,housing_df$Sale_Price, fill = housing_df$ctyname))

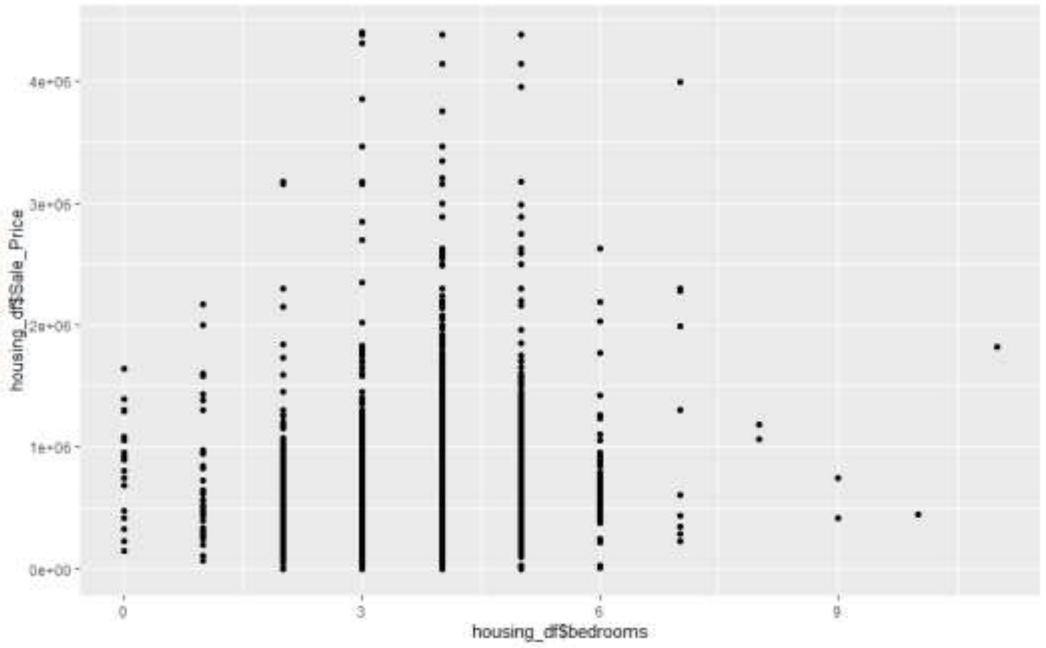
bar + stat_summary(fun = mean, geom = "bar", position="dodge",width = 8)+ facet_wrap(~ housing_df$ctyname)



#Sale_price are more in SAMMAMISH then REDMOND.

#There are some zips codes for which city name is NA.

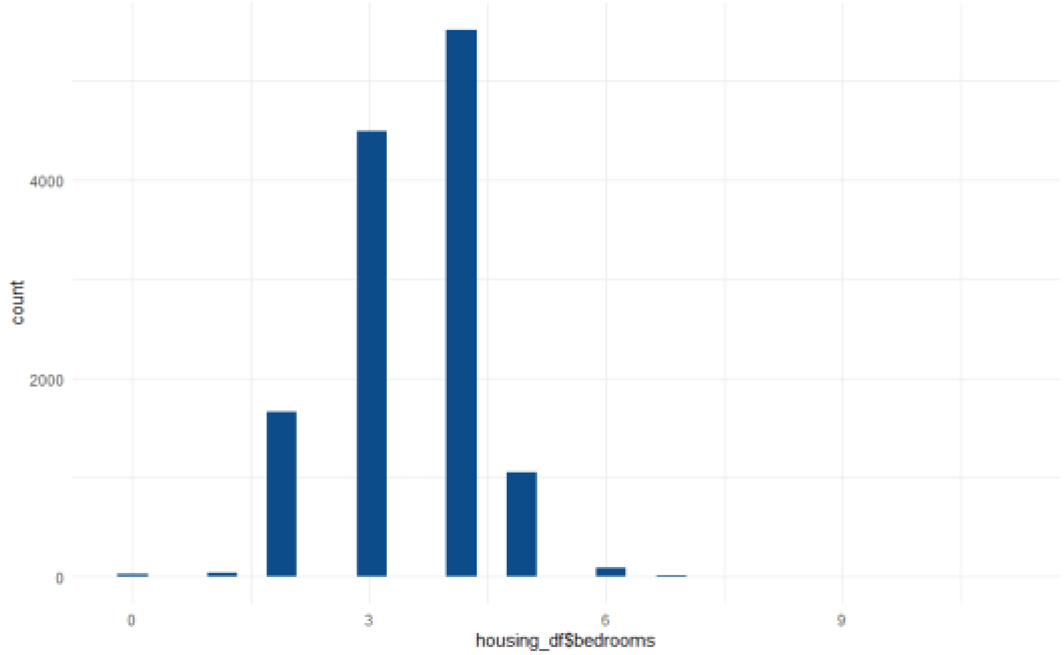
```
ggplot(housing_df, aes(x=housing_df$bedrooms, y=housing_df$Sale_Price)) + geom_point() +  xlim(0,
11)
```



```
#It looks like 4-bedroom houses are more popular for sale.
```

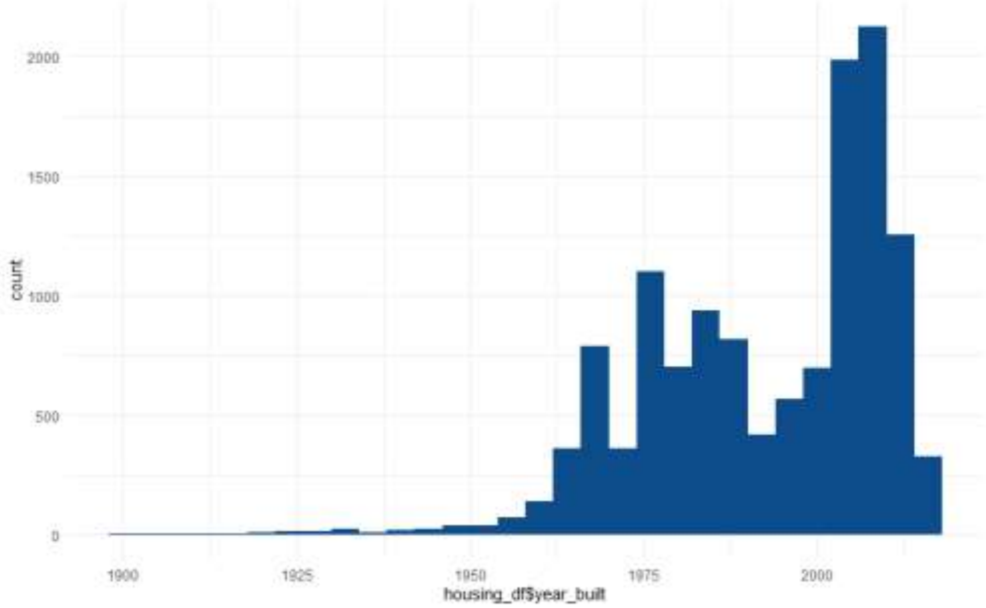
```
# e.Identify if there are any outliers
```

```
ggplot(housing_df) +
  aes(x = housing_df$bedrooms) +
  geom_histogram(bins = 30L, fill = "#0c4c8a") +
  theme_minimal()
```



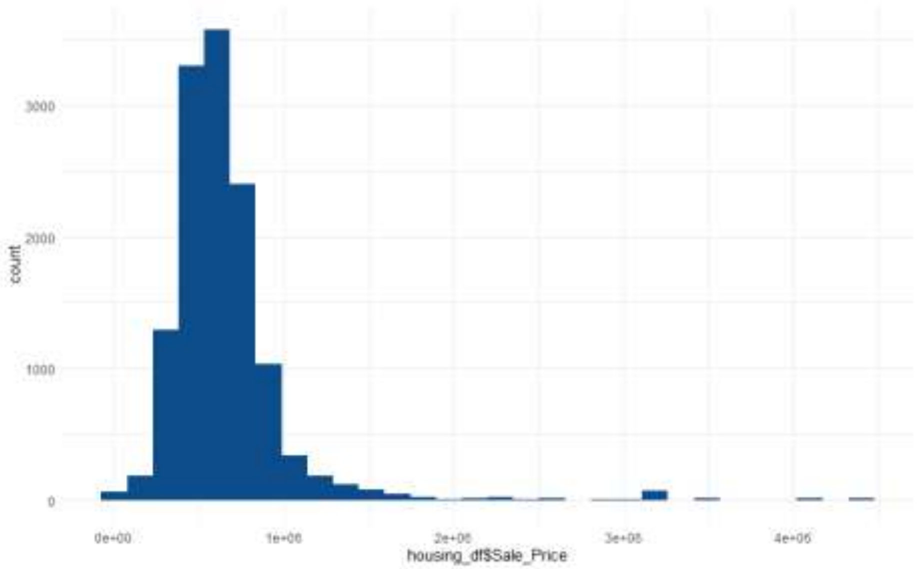
#All houses with bedroom >6 and <2 are outliers

```
ggplot(housing_df) +
  aes(x = housing_df$year_built) +
  geom_histogram(bins = 30L, fill = "#0c4c8a") +
  theme_minimal()
```



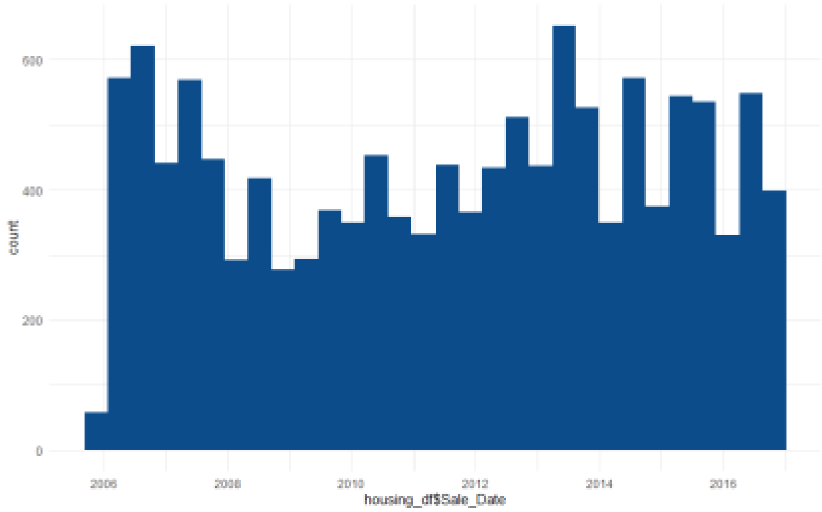
#All houses with built year < 1950 are outliers

```
ggplot(housing_df) +
  aes(x = housing_df$Sale_Price) +
  geom_histogram(bins = 30L, fill = "#0c4c8a") +
  theme_minimal()
```



#All houses with sales price > 2000000 are outliers

```
ggplot(housing_df) +
  aes(x = housing_df$Sale_Date) +
  geom_histogram(bins = 30L, fill = "#0c4c8a") +
  theme_minimal()
```



#There are no outliers for sale date

# f.Create at least 2 new variables

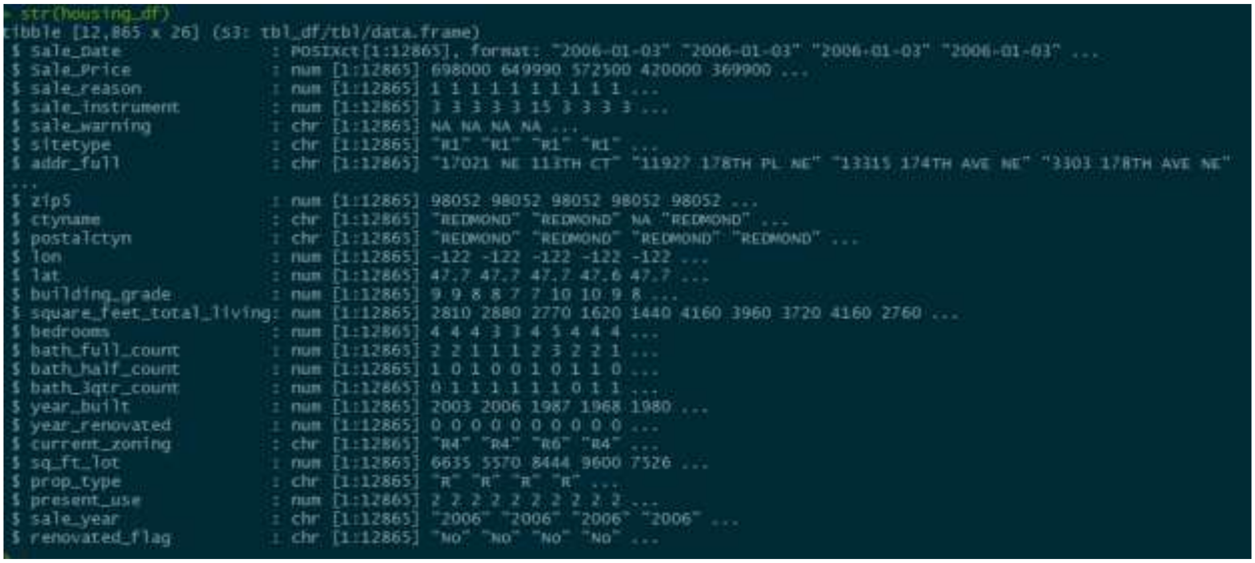
# deriving year of sale of the house

housing_df["sale_year"] <- substr(housing_df$Sale_Date,1,4)

# derive renovated flag

housing_df["renovated_flag"] <- ifelse(housing_df$year_renovated != 0, 'Yes', 'No')

str(housing_df)



```
> 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" "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 ...
 $ sale_year               : chr [1:12865] "2006" "2006" "2006" "2006" ...
 $ renovated_flag          : chr [1:12865] "No" "No" "No" "No" ...
```