DSC 520 - Assignment 4.2.2
Supraja Rapuru

```
> library(readxl)
> library(plyr)
> setwd("/Users/Supraja/dsc520")
> housing_df <- read_excel("data/week-6-housing.xlsx")
> 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>
> 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 ...
> apply(housing_df[,2],MARGIN=2,FUN=sum, na.rm=TRUE)
Sale Price
8500391149
> apply(housing_df[,1],MARGIN=2,FUN=sum, na.rm=TRUE)
Error in FUN(newX[, i], ...) : invalid 'type' (character) of argument
> colnames(housing_df)[1] <- "Sale_Date"
```

```
> colnames(housing_df)[2] <- "Sale_Price"
> aggregate(Sale_Price ~ ctyname, housing_df, mean)
   ctyname Sale_Price
1   REDMOND   644803.2
2 SAMMAMISH   972480.3
> 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
> summary(housing_df)
  Sale_Date                Sale_Price      sale_reason
 Min.   :2006-01-03 00:00:00   Min.   :    698   Min.   : 0.00
 1st Qu.:2008-07-07 00:00:00   1st Qu.: 460000   1st Qu.: 1.00
 Median :2011-11-17 00:00:00   Median : 593000   Median : 1.00
 Mean   :2011-07-28 15:07:32   Mean   : 660738   Mean   : 1.55
 3rd Qu.:2014-06-05 00:00:00   3rd Qu.: 750000   3rd Qu.: 1.00
 Max.   :2016-12-16 00:00:00   Max.   :4400000   Max.   :19.00
 sale_instrument  sale_warning       sitetype         addr_full
 Min.   : 0.000   Length:12865      Length:12865      Length:12865
 1st Qu.: 3.000   Class :character   Class :character   Class :character
 Median : 3.000   Mode  :character   Mode  :character   Mode  :character
 Mean   : 3.678
 3rd Qu.: 3.000
 Max.   :27.000
    zip5        ctyname          postalctyn           lon
 Min.   :98052   Length:12865      Length:12865      Min.   :-122.2
 1st Qu.:98052   Class :character   Class :character   1st Qu.:-122.1
 Median :98052   Mode  :character   Mode  :character   Median :-122.1
 Mean   :98053                                        Mean   :-122.1
 3rd Qu.:98053                                        3rd Qu.:-122.0
 Max.   :98074                                        Max.   :-121.9
    lat       building_grade  square_feet_total_living    bedrooms
 Min.   :47.46   Min.   : 2.00   Min.   :  240          Min.   : 0.000
 1st Qu.:47.67   1st Qu.: 8.00   1st Qu.: 1820          1st Qu.: 3.000
 Median :47.69   Median : 8.00   Median : 2420          Median : 4.000
 Mean   :47.68   Mean   : 8.24   Mean   : 2540          Mean   : 3.479
 3rd Qu.:47.70   3rd Qu.: 9.00   3rd Qu.: 3110          3rd Qu.: 4.000
 Max.   :47.73   Max.   :13.00   Max.   :13540          Max.   :11.000
 bath_full_count  bath_half_count  bath_3qtr_count   year_built
 Min.   : 0.000   Min.   :0.0000   Min.   :0.000   Min.   :1900
 1st Qu.: 1.000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1979
 Median : 2.000   Median :1.0000   Median :0.000   Median :1998
 Mean   : 1.798   Mean   :0.6134   Mean   :0.494   Mean   :1993
 3rd Qu.: 2.000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.:2007
 Max.   :23.000   Max.   :8.0000   Max.   :8.000   Max.   :2016
 year_renovated    current_zoning      sq_ft_lot       prop_type
 Min.   :  0.00   Length:12865      Min.   :    785   Length:12865
```

```
 1st Qu.:   0.00   Class :character   1st Qu.:   5355   Class :character
 Median :   0.00   Mode  :character   Median :   7965   Mode  :character
 Mean   :  26.24                      Mean   :  22229
 3rd Qu.:   0.00                      3rd Qu.:  12632
 Max.   :2016.00                      Max.   :1631322
  present_use
 Min.   :  0.000
 1st Qu.:  2.000
 Median :  2.000
 Mean   :  6.598
 3rd Qu.:  2.000
 Max.   :300.000
> unique(housing_df$prop_type)
[1] "R"
> unique(housing_df$ctyname)
[1] "REDMOND"   NA          "SAMMAMISH"
> library(ggplot2)
> bar + stat_summary(fun = mean, geom = "bar", position="dodge",width = 8)+ facet_wrap( ~
housing_df$ctyname)
Error in `$<-.data.frame`(`*tmp*`, "PANEL", value = c(1L, 1L, 3L, 1L,  :
  replacement has 12865 rows, data has 38
> 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)
Warning messages:
1: Use of `housing_df$zip5` is discouraged. Use `zip5` instead.
2: Use of `housing_df$Sale_Price` is discouraged. Use `Sale_Price` instead.
3: Use of `housing_df$ctyname` is discouraged. Use `ctyname` instead.
4: position_dodge requires non-overlapping x intervals
> ggplot(housing_df, aes(x=housing_df$bedrooms, y=housing_df$Sale_Price)) + geom_point() +
xlim(0, 11)
Warning messages:
1: Use of `housing_df$bedrooms` is discouraged. Use `bedrooms` instead.
2: Use of `housing_df$Sale_Price` is discouraged. Use `Sale_Price` instead.
> ggplot(housing_df) +
+   aes(x = housing_df$bedrooms) +
+   geom_histogram(bins = 30L, fill = "#0c4c8a") +
+   theme_minimal()
Warning message:
Use of `housing_df$bedrooms` is discouraged. Use `bedrooms` instead.
> ggplot(housing_df) +
+   aes(x = housing_df$year_built) +
+   geom_histogram(bins = 30L, fill = "#0c4c8a") +
+   theme_minimal()
Warning message:
Use of `housing_df$year_built` is discouraged. Use `year_built` instead.
> ggplot(housing_df) +
+   aes(x = housing_df$Sale_Price) +
+   geom_histogram(bins = 30L, fill = "#0c4c8a") +
+   theme_minimal()
Warning message:
Use of `housing_df$Sale_Price` is discouraged. Use `Sale_Price` instead.
> ggplot(housing_df) +
+   aes(x = housing_df$Sale_Date) +
+   geom_histogram(bins = 30L, fill = "#0c4c8a") +
+   theme_minimal()
Warning message:
```

```
Use of `housing_df$Sale_Date` is discouraged. Use `Sale_Date` instead.
> housing_df["sale_year"] <- substr(housing_df$Sale_Date,1,4)
> housing_df["renovated_flag"] <- ifelse(housing_df$year_renovated != 0, 'Yes', 'No')
> 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 ...
 $ sale_year             : chr [1:12865] "2006" "2006" "2006" "2006" ...
 $ renovated_flag        : chr [1:12865] "No" "No" "No" "No" ...
>
```