DSC520_Week4_Assignment01

Reenie Christudass

2022-07-11

```
library(crayon)
```

```
## Warning: package 'crayon' was built under R version 4.2.1
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.2.1
```

```
##
## 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(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 4.2.1
```

```
library(readxl)
df <- read_excel("C:/Users/chris/dsc520/data/week-7-housing.xlsx")
cat(blue("Read the Week 7 Housing excel document\n"))
```

```
## Read the Week 7 Housing excel document
```

```
## Summary of each column
cat(blue("Summary of variables in the dataset\n"))
```

```
## Summary of variables in the dataset
```

```
summary(df)
```

```
##    Sale Date                        Sale Price       sale_reason
##  Min.   :2006-01-03 00:00:00.00   Min.   :    698   Min.   : 0.00
##  1st Qu.:2008-07-07 00:00:00.00   1st Qu.: 460000   1st Qu.: 1.00
##  Median :2011-11-17 00:00:00.00   Median : 593000   Median : 1.00
##  Mean   :2011-07-28 15:07:32.48   Mean   : 660738   Mean   : 1.55
##  3rd Qu.:2014-06-05 00:00:00.00   3rd Qu.: 750000   3rd Qu.: 1.00
##  Max.   :2016-12-16 00:00:00.00   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
```

```
##Remove space in the column name
cat(blue("Remove space in the column name"  ))
```

```
## Remove space in the column name
```

```
names(df) <- sub(" ", "_", names(df))
head(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>
```

```
##Use the aggregate function on a variable in your dataset
group_mean <- aggregate(Sale_Price ~ year_built   , data = df, FUN = sum)
head(group_mean)
```

```
##   year_built Sale_Price
## 1       1900    2366998
## 2       1903     430000
## 3       1905     620000
## 4       1906     550000
## 5       1909       1070
## 6       1910     150000
```

```
cat(blue("Seperate the Sale_Date into three variables Year, Month, Date\n"))
```

```
## Seperate the Sale_Date into three variables Year, Month, Date
```

```
df <- df %>% separate(Sale_Date, c('Year', 'Month','Date'))
head(df)
```

```
## # A tibble: 6 x 26
##   Year  Month Date  Sale_Price sale_reason sale_instrument sale_warning sitetype
##   <chr> <chr> <chr>      <dbl>       <dbl>           <dbl> <chr>        <chr>
## 1 2006  01    03        698000           1               3 <NA>         R1
## 2 2006  01    03        649990           1               3 <NA>         R1
## 3 2006  01    03        572500           1               3 <NA>         R1
## 4 2006  01    03        420000           1               3 <NA>         R1
## 5 2006  01    03        369900           1               3 15           R1
## 6 2006  01    03        184667           1              15 18 51        R1
## # ... with 18 more variables: 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>
```

```
## Re-Create the Column Sale_price
cat(blue("Re-Create the Column Sale_price\n"))
```

```
## Re-Create the Column Sale_price
```

```
df$Sale_Date <- paste(df$Year,"-",df$Month,"-",df$Date)
head(df)
```

```
## # A tibble: 6 x 27
##   Year  Month Date  Sale_Price sale_reason sale_instrument sale_warning sitetype
##   <chr> <chr> <chr>      <dbl>       <dbl>           <dbl> <chr>        <chr>
## 1 2006  01    03        698000           1               3 <NA>         R1
## 2 2006  01    03        649990           1               3 <NA>         R1
## 3 2006  01    03        572500           1               3 <NA>         R1
## 4 2006  01    03        420000           1               3 <NA>         R1
## 5 2006  01    03        369900           1               3 15           R1
## 6 2006  01    03        184667           1              15 18 51        R1
## # ... with 19 more variables: 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>, Sale_Date <chr>
```

```
##Re-locate the Column Sale_Date in front of Sale_Price
cat(blue("Re-locate the Column Sale_Date in front of Sale_Price\n"))
```

```
## Re-locate the Column Sale_Date in front of Sale_Price
```

```
df %>% relocate(Sale_Date, .before = Sale_Price)
```

```
## # A tibble: 12,865 x 27
##    Year  Month Date  Sale_Date      Sale_Price sale_reason sale_instrument
##    <chr> <chr> <chr> <chr>               <dbl>       <dbl>           <dbl>
##  1 2006  01    03    2006 - 01 - 03     698000           1               3
##  2 2006  01    03    2006 - 01 - 03     649990           1               3
##  3 2006  01    03    2006 - 01 - 03     572500           1               3
##  4 2006  01    03    2006 - 01 - 03     420000           1               3
##  5 2006  01    03    2006 - 01 - 03     369900           1               3
##  6 2006  01    03    2006 - 01 - 03     184667           1              15
##  7 2006  01    04    2006 - 01 - 04    1050000           1               3
##  8 2006  01    04    2006 - 01 - 04     875000           1               3
##  9 2006  01    04    2006 - 01 - 04     660000           1               3
## 10 2006  01    04    2006 - 01 - 04     650000           1               3
## # ... with 12,855 more rows, and 20 more variables: sale_warning <chr>,
## #   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>
```

```
##perform a modification to the data, and then bring it back together
df <-dplyr::select(df, -c('Year', 'Month','Date'))
df
```

```
## # A tibble: 12,865 x 24
##    Sale_Price sale_reason sale_instrument sale_warning sitetype addr_full   zip5
##         <dbl>       <dbl>           <dbl> <chr>        <chr>    <chr>      <dbl>
##  1     698000           1               3 <NA>         R1       17021 NE ~ 98052
##  2     649990           1               3 <NA>         R1       11927 178~ 98052
##  3     572500           1               3 <NA>         R1       13315 174~ 98052
##  4     420000           1               3 <NA>         R1       3303 178T~ 98052
##  5     369900           1               3 15           R1       16126 NE ~ 98052
##  6     184667           1              15 18 51        R1       8101 229T~ 98053
##  7    1050000           1               3 <NA>         R1       21634 NE ~ 98053
##  8     875000           1               3 <NA>         R1       21404 NE ~ 98053
##  9     660000           1               3 <NA>         R1       7525 238T~ 98053
## 10     650000           1               3 <NA>         R1       17703 NE ~ 98052
## # ... with 12,855 more rows, and 17 more variables: 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>, Sale_Date <chr>
```

```
## Identify if there are any outliers
cat(blue("Mean of the variable bedrooms\n"))
```

```
## Mean of the variable bedrooms
```

```
mean(df$bedrooms)
```

```
## [1] 3.478663
```

```
cat(blue("Median of the variable bedrooms\n"))
```

```
## Median of the variable bedrooms
```

```
median(df$bedrooms)
```

```
## [1] 4
```

```
cat(blue("Quantile of the variable bedrooms\n"))
```

```
## Quantile of the variable bedrooms
```

```
quantile(df$bedrooms)
```

```
##   0%  25%  50%  75% 100%
##    0    3    4    4   11
```

```
cat(blue("Outliers of the variable bedrooms\n"))
```

```
## Outliers of the variable bedrooms
```

```
boxplot(df$bedrooms,
  ylab = "bedrooms",
  main = "Boxplot of bedrooms \n"
)
```

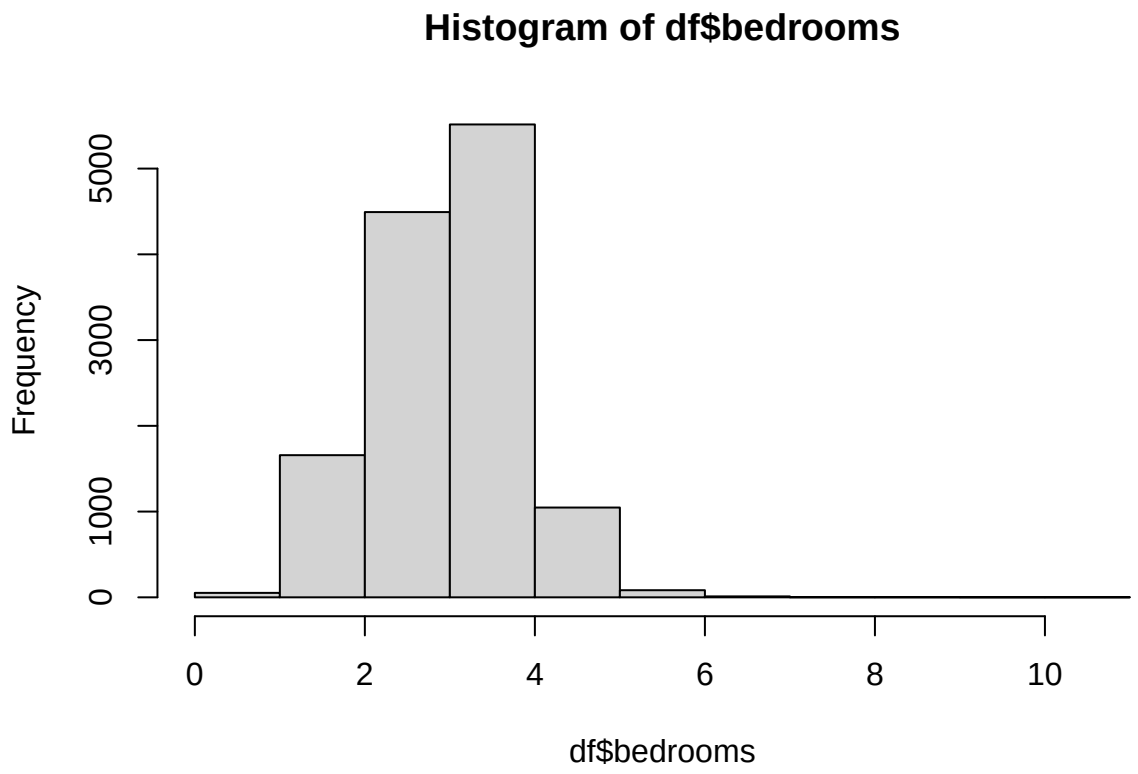

```
cat(blue("Outliers of the variable bedrooms\n"))
```

```
## Outliers of the variable bedrooms
```

```
out <- boxplot.stats(df$bedrooms)$out
out
```

```
##   [1]  6  0  6  0  6  6  6  6  6  9  6  7  6  1  0  8  6  6  0  6  6  6  7  6  0
##  [26]  1  1  6  1  6  1  0  6 10  0  1  7  6 11  6  6  6  6  7  7  6  6  6  6  6
##  [51]  6  6  6  6  1  1  6  0  1  7  6  1  1  6  1  0  0  1  6  1  6  0  6  1  1
##  [76]  6  6  6  7  1  6  1  0  6  6  6  7  6  1  6  6  7  1  7  1  1  6  0  6  1
## [101]  9  6  6  6  6  7  1  6  0  6  6  6  6  6  1  1  6  1  6  6  6  6  6  0  8
## [126]  6  6  6  6  6  6  6  6  1  0  1  6  6  6  0  1  0  6  6  1  6  6  6  1  1
## [151]  6  0
```

```
hist(df$bedrooms)
```




```
cat(blue("Summary of the variable bedrooms\n"))
```

```
## Summary of the variable bedrooms
```

```
summary(df$bedrooms)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   3.000   4.000   3.479   4.000  11.000
```