BUSN 32100 HW6

Mia Jiang

11/15/2022

1. Setting WD and Loading Libraries

```
### Clear Global Environment
knitr::opts_chunk$set(
  echo = TRUE,
  message = FALSE,
  warning = FALSE,
  fig.height = 4,
  fig.width = 6
)
rm(list = ls())
options(scipen = 0)
# Setting the Working Directory
setwd("~/Desktop/Fall Quarter/BUSN-32100/Week8/BUSN-32100-HW-6")
# Load packages
library(readr) # read csv
library(dplyr) # data manipulation
library(tidyverse)
library(caret) # print text
library(reshape2) # melt data for correlation map plotting
library(ggplot2) # plotting
library(GGally) # pair plot
library(plotly) # interactive plot
```

2. Import data

```
# Load data
kc_housing <- read_csv("kc_house_data.csv", show_col_types = FALSE)
```

3. Summary Statistics

```
as.data.frame(summary(kc_housing)) %>%
  select(-Var1) %>%
  separate(Freq, into = c("attribute", "value"), sep = ":", extra = "merge") %>%
  mutate(attribute = str_remove(attribute, "\\.")) %>%
  pivot_wider(names_from = Var2,
              values_from = value)
```

```
## # A tibble: 6 x 22
##   attribute `      id`    `     date`    `    price` `   bedrooms` `  bathrooms`
##   <chr>     <chr>         <chr>          <chr>       <chr>         <chr>
## 1 "Min   "  "1.000e+06  " "2014-05-02 0~ "  75000  " " 0.000  "    "0.000  "
## 2 "1st Qu"  "2.123e+09  " "2014-07-22 0~ " 321950  " " 3.000  "    "1.750  "
## 3 "Median " "3.905e+09  " "2014-10-16 0~ " 450000  " " 3.000  "    "2.250  "
## 4 "Mean   " "4.580e+09  " "2014-10-29 0~ " 540088  " " 3.371  "    "2.115  "
## 5 "3rd Qu"  "7.309e+09  " "2015-02-17 0~ " 645000  " " 4.000  "    "2.500  "
## 6 "Max   "  "9.900e+09  " "2015-05-27 0~ "7700000  " "33.000  "    "8.000  "
## # ... with 16 more variables: ` sqft_living` <chr>, `   sqft_lot` <chr>,
## #   `    floors` <chr>, `  waterfront` <chr>, `     view` <chr>,
## #   `  condition` <chr>, `    grade` <chr>, `  sqft_above` <chr>,
## #   sqft_basement <chr>, `   yr_built` <chr>, ` yr_renovated` <chr>,
## #   `   zipcode` <chr>, `     lat` <chr>, `     long` <chr>,
## #   sqft_living15 <chr>, `  sqft_lot15` <chr>
```

4. Information on All Columns

```
# Column info
glimpse(kc_housing)
```

```
## Rows: 21,613
## Columns: 21
## $ id            <dbl> 7129300520, 6414100192, 5631500400, 2487200875, 19544005~
## $ date          <dttm> 2014-10-13, 2014-12-09, 2015-02-25, 2014-12-09, 2015-02~
## $ price         <dbl> 221900, 538000, 180000, 604000, 510000, 1225000, 257500,~
## $ bedrooms      <dbl> 3, 3, 2, 4, 3, 4, 3, 3, 3, 3, 3, 2, 3, 3, 5, 4, 3, 4, 2,~
## $ bathrooms     <dbl> 1.00, 2.25, 1.00, 3.00, 2.00, 4.50, 2.25, 1.50, 1.00, 2.~
## $ sqft_living   <dbl> 1180, 2570, 770, 1960, 1680, 5420, 1715, 1060, 1780, 189~
## $ sqft_lot      <dbl> 5650, 7242, 10000, 5000, 8080, 101930, 6819, 9711, 7470,~
## $ floors        <dbl> 1.0, 2.0, 1.0, 1.0, 1.0, 1.0, 2.0, 1.0, 1.0, 2.0, 1.0, 1~
## $ waterfront    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ view          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3, 0, 0, 0,~
## $ condition     <dbl> 3, 3, 3, 5, 3, 3, 3, 3, 3, 3, 3, 4, 4, 4, 3, 3, 3, 4, 4,~
## $ grade         <dbl> 7, 7, 6, 7, 8, 11, 7, 7, 7, 7, 8, 7, 7, 7, 7, 9, 7, 7, 7~
## $ sqft_above    <dbl> 1180, 2170, 770, 1050, 1680, 3890, 1715, 1060, 1050, 189~
## $ sqft_basement <dbl> 0, 400, 0, 910, 0, 1530, 0, 0, 730, 0, 1700, 300, 0, 0, ~
## $ yr_built      <dbl> 1955, 1951, 1933, 1965, 1987, 2001, 1995, 1963, 1960, 20~
## $ yr_renovated  <dbl> 0, 1991, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ zipcode       <dbl> 98178, 98125, 98028, 98136, 98074, 98053, 98003, 98198, ~
## $ lat           <dbl> 47.5112, 47.7210, 47.7379, 47.5208, 47.6168, 47.6561, 47~
## $ long          <dbl> -122.257, -122.319, -122.233, -122.393, -122.045, -122.0~
## $ sqft_living15 <dbl> 1340, 1690, 2720, 1360, 1800, 4760, 2238, 1650, 1780, 23~
## $ sqft_lot15    <dbl> 5650, 7639, 8062, 5000, 7503, 101930, 6819, 9711, 8113, ~
```

```
# Missing values
summary(colSums(is.na(kc_housing)))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       0       0       0       0       0       0
```

Answer

- There are 21513 observations with 21 variables in total. For all columns, no missing value has been found in this data set.

5. Rename a column

```
# Rename columns
kc_housing <- kc_housing %>%
  rename("sold_price" = "price", "constr_quality" = "grade")
# Check column names
names(kc_housing)
```

```
##  [1] "id"             "date"           "sold_price"     "bedrooms"
##  [5] "bathrooms"      "sqft_living"    "sqft_lot"       "floors"
##  [9] "waterfront"     "view"           "condition"      "constr_quality"
## [13] "sqft_above"     "sqft_basement"  "yr_built"       "yr_renovated"
## [17] "zipcode"        "lat"            "long"           "sqft_living15"
## [21] "sqft_lot15"
```

Note

- Renaming the *price* column to make it more clear that the price refers to *sold price*.
- Renaming the *grade* column to *constr_quality* so that it's more clear this column records **construction quality grade** of houses.

6. Count number of houses by bedroom

```
count(kc_housing, bedrooms) %>%
  arrange(desc(n))
```

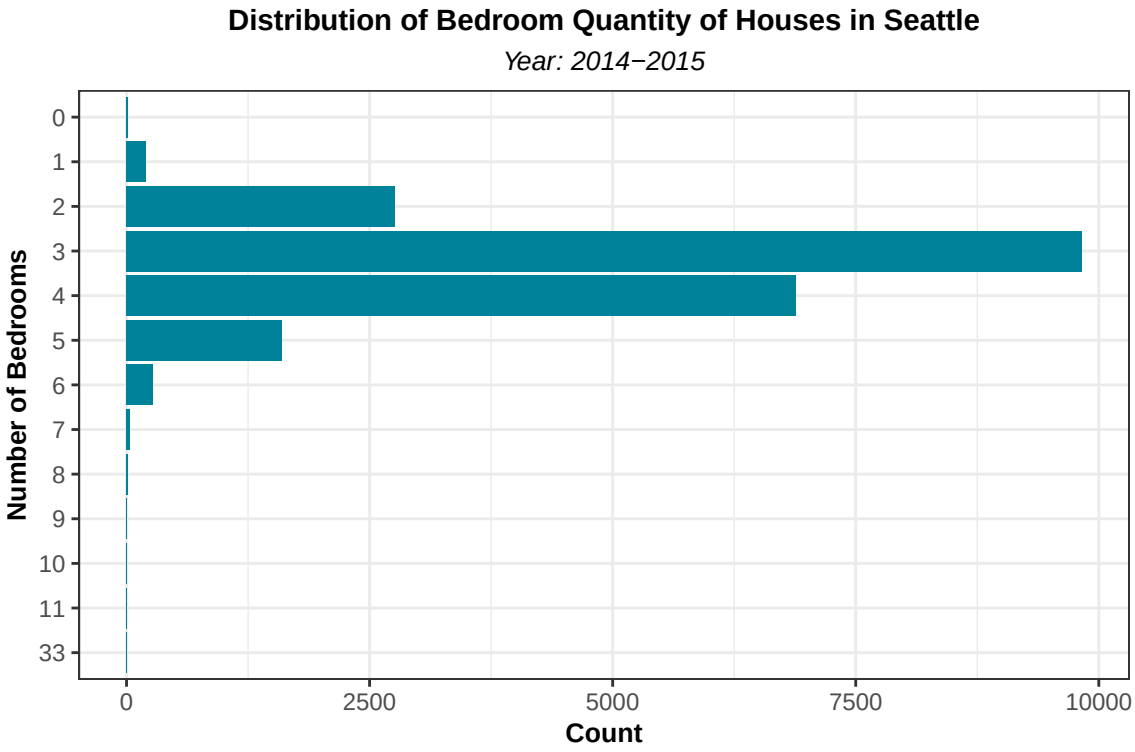
```
## # A tibble: 13 x 2
##    bedrooms     n
##       <dbl> <int>
##  1        3  9824
##  2        4  6882
##  3        2  2760
##  4        5  1601
##  5        6   272
##  6        1   199
##  7        7    38
##  8        0    13
##  9        8    13
## 10        9     6
## 11       10     3
## 12       11     1
## 13       33     1
```

Answer

- 3-bedroom is the most popular type, as nearly half of the houses in this dataset have 3 bedrooms, followed by 4-bedroom and 2-bedroom types. In general, 3 or 4 bedrooms are the most common ones in Seattle, and it's relatively rare that a house has 5+ or less than 2 bedrooms.

7. Bar Chart of the Number of Bedrooms in Seattle Homes

```
# Set color and theme
seattle <- "#00839A"
theme <- theme_bw() +
  theme(
    plot.title = element_text(size = 11, face = "bold", hjust = 0.5),
    plot.subtitle = element_text(size = 10, face = "italic", hjust = 0.5),
    axis.title = element_text(size = 10, face = "bold")
  )
# Plot
ggplot(kc_housing) +
  geom_bar(aes(
    x = fct_reorder(as.factor(bedrooms), desc(bedrooms))
  ),
  stat = "count",
  fill = seattle
  ) +
  labs(
    title = "Distribution of Bedroom Quantity of Houses in Seattle",
    subtitle = "Year: 2014-2015",
    x = "Number of Bedrooms",
    y = "Count"
  ) +
  coord_flip() +
  theme
```



8. Mean Number of Bedrooms

```
cat("The mean number of bedrooms in seattle houses between 2014 to 2015 is",
      mean(kc_housing$bedrooms))
```

```
## The mean number of bedrooms in seattle houses between 2014 to 2015 is 3.370842
```

9. Subsetting and selecting rows

```
# Filter on bedrooms less than 5
clean_kc_housing <- kc_housing %>%
  filter(bedrooms < 5)
# Check new data
cat(
  "After excluding houses with 5 or more bedrooms, there are",
  nrow(clean_kc_housing), "observations left, which means",
  str_c(format(nrow(clean_kc_housing) / nrow(kc_housing) * 100, digits = 2), "%"),
  "houses in Seattle have less than 5 bedrooms between Year 2014 and 2015.",
  "Houses with 5 or more bedrooms are not typical."
)
```

```
## After excluding houses with 5 or more bedrooms, there are 19678 observations left, which means 91% ho
```

10. Filter on multiple criteria

```
# Filter on bedrooms less than 3 and living footage greater than 500
clean_kc_housing_2 <- kc_housing %>%
  filter(bedrooms < 3 & sqft_living > 500)
# Check new data
nrow(clean_kc_housing_2)
```

```
## [1] 2955
```

```
# Filter on bedrooms greater than 3 or floors greater than 2
clean_kc_housing_3 <- kc_housing %>%
  filter(bedrooms > 3 | floors > 2)
# Check new data
nrow(clean_kc_housing_3)
```

```
## [1] 9419
```

11. Find null values for price_sold

```
summary(is.na(kc_housing$sold_price))
```

```
##    Mode   FALSE
## logical   21613
```

```
summary(kc_housing$sold_price == "")
```

```
##    Mode   FALSE
## logical   21613
```

Answer

- There is no null value for the sold price of houses in this data set.

12. Group by and Apply on a Column

```
kc_housing %>%
  group_by(view) %>% # group by view score
  summarise(mean_price = mean(sold_price), # calculate mean sold price
            median_price = median(sold_price)) # calculate median sold price
```

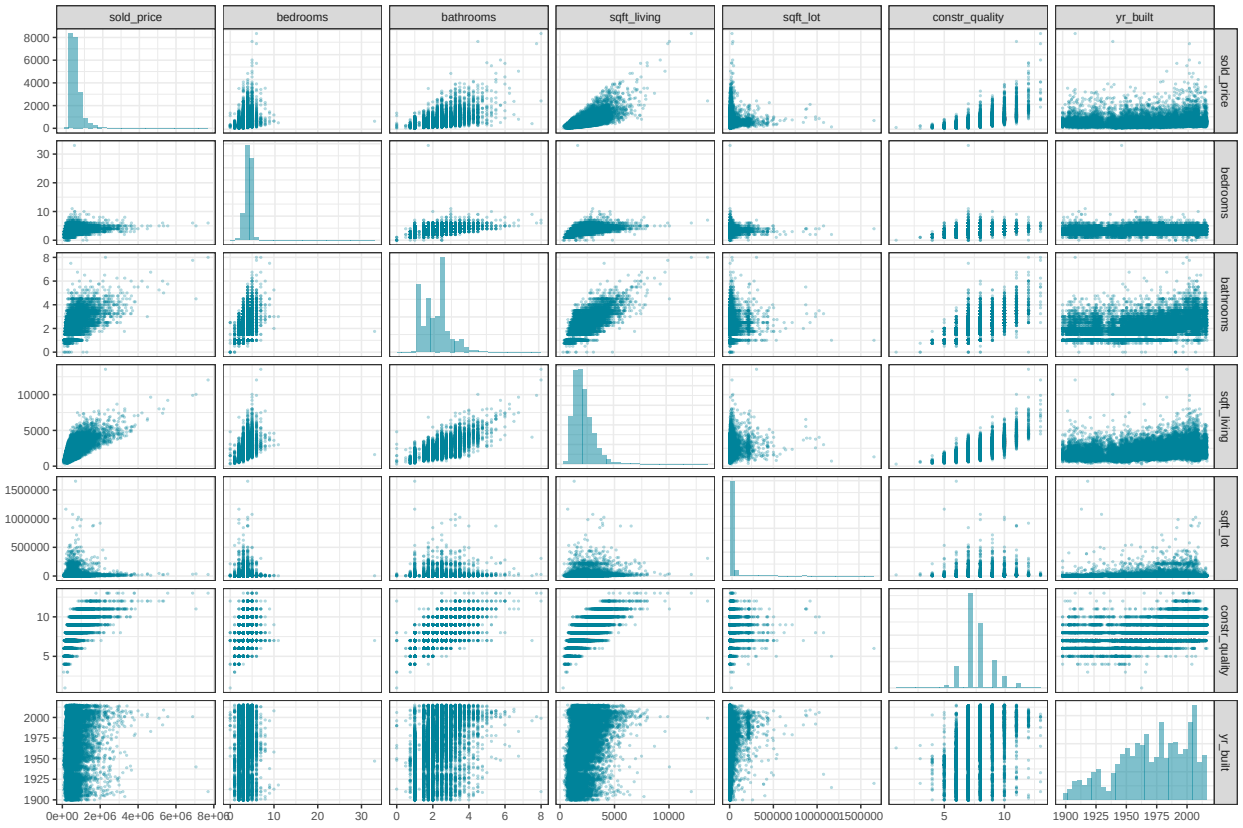
```
## # A tibble: 5 x 3
##    view mean_price median_price
##   <dbl>      <dbl>        <dbl>
## 1     0    496564.       432500
## 2     1    812281.       690944
## 3     2    792401.       675000
## 4     3    971965.       802500
## 5     4   1463711.      1185000
```

Answer

- Unsurprisingly, houses with better views were sold with higher mean and median prices, i.e.e they are more expensive.

13. Pairplot of Numeric Columns

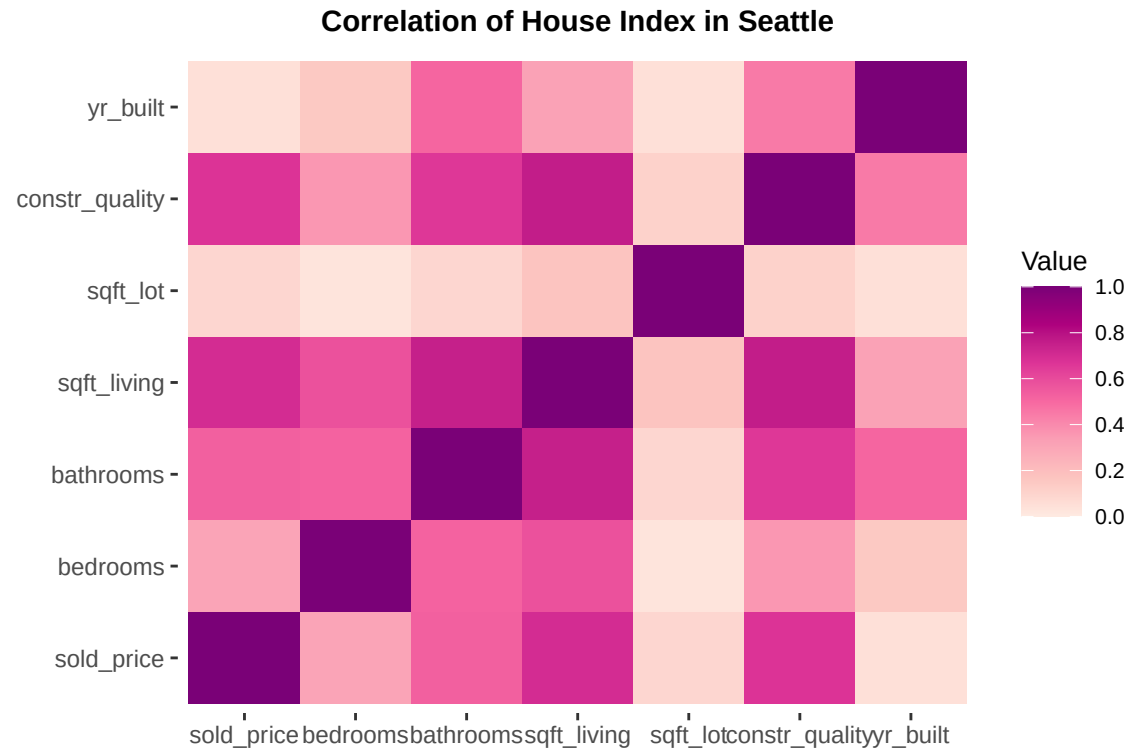
```
# Select informative numeric columns
kc_housing_num <- kc_housing %>%
  select(sold_price:sqft_lot, constr_quality, yr_built)
# Plot
my_facet <- function(data, mapping, ...) {
  ggplot(data, mapping) +
    geom_point(...,
               color = seattle, alpha = 0.3, size = 0.1)
}

# Plot
ggpairs(kc_housing_num,
  lower = list(continuous = my_facet),
  upper = list(continuous = my_facet),
  diag = list(continuous = wrap("barDiag", bins = 30,
                                alpha = 0.5, fill = seattle))
  ) +
  theme_bw(base_size = 8) +
  theme(axis.text = element_text(size = 6))
```



14. Correlation Heatmap

```
# Calculate correlation
corr_mat <- round(cor(kc_housing_num), 2)
# Reshape correlation data
melted_corr_mat <- melt(corr_mat)
# Plot
plt <- ggplot(melted_corr_mat) +
  geom_tile(aes(Var1, Var2, fill = value)) +
  scale_fill_distiller(
    name = "Value",
    palette = "RdPu",
    direction = 1,
    breaks = seq(0, 1, 0.2),
    limits = c(0,1)
  ) +
  labs(title = "Correlation of House Index in Seattle") +
  theme(
    plot.title = element_text(size = 11, face = "bold", hjust = 0.5),
    axis.title = element_blank(),
    legend.title = element_text(size = 10),
    legend.text = element_text(size = 8),
    legend.key.width = unit(0.8, "cm"),
    panel.background = element_blank()
  )
```

```
plt
```




```
# ggplotly(plt)
```