FA24 Lab 6 In Class Exercise

Solutions to the Class Exercise

1. Install Packages

```
rm(list = ls())
library(pacman)
p_load(ggplot2, dplyr, modelsummary)
```

2. Load Data

```
txhousing <- ggplot2::txhousing %>% tibble()
```

3. Data Exploration

```
txhousing %>% summary()
```

```
##      city                year          month            sales       
##  Length:8602        Min.   :2000   Min.   : 1.000   Min.   :   6.0  
##  Class :character   1st Qu.:2003   1st Qu.: 3.000   1st Qu.:  86.0  
##  Mode  :character   Median :2007   Median : 6.000   Median : 169.0  
##                     Mean   :2007   Mean   : 6.406   Mean   : 549.6  
##                     3rd Qu.:2011   3rd Qu.: 9.000   3rd Qu.: 467.0  
##                     Max.   :2015   Max.   :12.000   Max.   :8945.0  
##                                                     NA's   :568     
##      volume             median          listings       inventory     
##  Min.   :8.350e+05   Min.   : 50000   Min.   :    0   Min.   : 0.000  
##  1st Qu.:1.084e+07   1st Qu.:100000   1st Qu.:  682   1st Qu.: 4.900  
##  Median :2.299e+07   Median :123800   Median : 1283   Median : 6.200  
##  Mean   :1.069e+08   Mean   :128131   Mean   : 3217   Mean   : 7.175  
##  3rd Qu.:7.512e+07   3rd Qu.:150000   3rd Qu.: 2954   3rd Qu.: 8.150  
##  Max.   :2.568e+09   Max.   :304200   Max.   :43107   Max.   :55.900  
##  NA's   :568         NA's   :616      NA's   :1424    NA's   :1467    
##       date        
##  Min.   :2000  
##  1st Qu.:2004  
##  Median :2008  
##  Mean   :2008  
##  3rd Qu.:2012  
##  Max.   :2016  
##
```

```
txhousing %>% str()
```

```
## tibble [8,602 x 9] (S3: tbl_df/tbl/data.frame)
##  $ city     : chr [1:8602] "Abilene" "Abilene" "Abilene" "Abilene" ...
##  $ year     : int [1:8602] 2000 2000 2000 2000 2000 2000 2000 2000 2000 2000 ...
##  $ month    : int [1:8602] 1 2 3 4 5 6 7 8 9 10 ...
##  $ sales    : num [1:8602] 72 98 130 98 141 156 152 131 104 101 ...
##  $ volume   : num [1:8602] 5380000 6505000 9285000 9730000 10590000 ...
##  $ median   : num [1:8602] 71400 58700 58100 68600 67300 66900 73500 75000 64500 59300 ...
##  $ listings : num [1:8602] 701 746 784 785 794 780 742 765 771 764 ...
##  $ inventory: num [1:8602] 6.3 6.6 6.8 6.9 6.8 6.6 6.2 6.4 6.5 6.6 ...
##  $ date     : num [1:8602] 2000 2000 2000 2000 2000 ...
```

4. Replace NA in variables

```
# Before
# Count missings manually
txhousing %>% mutate(sales = is.na(sales)) %>%
              mutate(volume = is.na(volume)) %>%
              mutate(median = is.na(median)) %>%
              mutate(listings = is.na(listings)) %>%
              mutate(inventory = is.na(inventory)) %>%
              summarize(sales     = sum(sales),
                        volume    = sum(volume),
                        median    = sum(median),
                        listings  = sum(listings),
                        inventory = sum(inventory)) %>% t()
```

```
##           [,1]
## sales      568
## volume     568
## median     616
## listings  1424
## inventory 1467
```

```
# After
txhousing_clean <- txhousing %>% replace(is.na(.),0)
# Count missings manually
txhousing_clean %>% mutate(sales = is.na(sales)) %>%
              mutate(volume = is.na(volume)) %>%
              mutate(median = is.na(median)) %>%
              mutate(listings = is.na(listings)) %>%
              mutate(inventory = is.na(inventory)) %>%
              summarize(sales     = sum(sales),
                        volume    = sum(volume),
                        median    = sum(median),
                        listings  = sum(listings),
                        inventory = sum(inventory)) %>% t()
```

```
##           [,1]
```

```
## sales        0
## volume       0
## median       0
## listings     0
## inventory    0
```

5. Summarize

```
txhousing_year <- txhousing_clean %>% dplyr::group_by(city,year) %>%
                                      dplyr::summarize(sales     = mean(sales, na.rm =TRUE),
                                                       volume    = mean(volume, na.rm =TRUE),
                                                       median    = mean(median, na.rm =TRUE),
                                                       listings  = mean(listings, na.rm =TRUE),
                                                       inventory = mean(inventory, na.rm =TRUE))
```

```
## `summarise()` has grouped output by 'city'. You can override using the
## `.groups` argument.
```

```
txhousing_year %>% head(n=10)
```

```
## # A tibble: 10 x 7
## # Groups:   city [1]
##    city     year sales    volume  median listings inventory
##    <chr>   <int> <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
##  1 Abilene  2000  115.  9047917.  66600      751.      6.47
##  2 Abilene  2001  119.  9530417.  70975      770.      6.62
##  3 Abilene  2002  126.  9889583.  68600      716.      5.84
##  4 Abilene  2003  136  11306250   71933.     726.      5.68
##  5 Abilene  2004  152. 13305833.  74867.     678.      4.56
##  6 Abilene  2005  165. 16571250   87592.     612       3.82
##  7 Abilene  2006  166. 18960833. 100292.     673.      4.11
##  8 Abilene  2007  167. 19338549. 103650      864.      4.96
##  9 Abilene  2008  138. 16043361. 107133.     924       6.32
## 10 Abilene  2009  136. 16863146. 108367.     812.      6.12
```

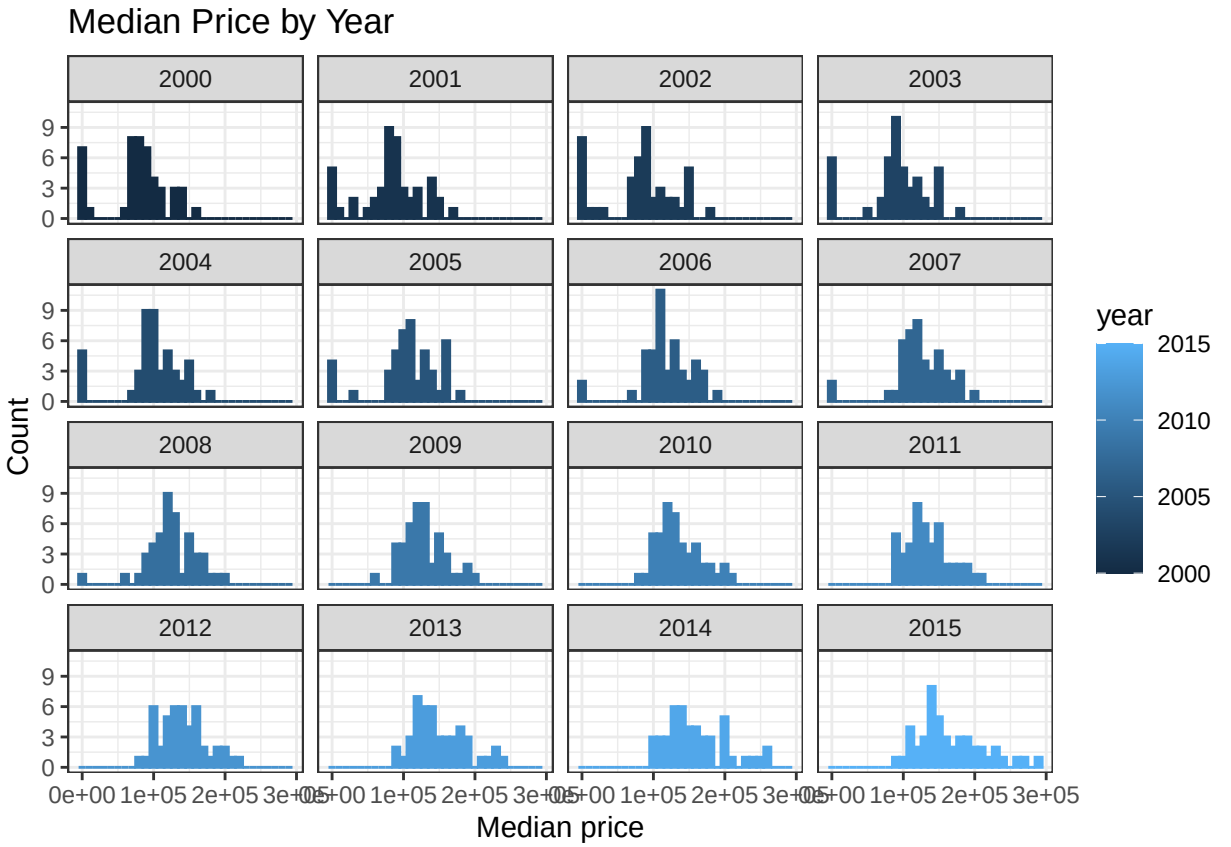
6. Scatterplot

```
txhousing_year %>% ggplot(mapping = aes(x = sales, y = median, color = year)) +
                   geom_point(alpha = 0.5) +
                   labs(x = "Sales", y = "Median Price", title = "Sales and Median Price") +
                   theme_bw()
```

7. Histogram

```
txhousing_year %>% ggplot(mapping = aes(x = median, fill = year, color = year)) +
                   geom_histogram() + facet_wrap(~year) +
                   labs(x = "Median price", y = "Count", title = "Median Price by Year") +
                   theme_bw()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



8. Filter by Year

```
txhousing_year07 <- txhousing_year %>% dplyr::filter(year >= 2007 & year <= 2009)
txhousing_year07 %>% head(n=10)
```

```
## # A tibble: 10 x 7
## # Groups:   city [4]
##    city       year sales     volume  median listings inventory
##    <chr>     <int> <dbl>      <dbl>   <dbl>    <dbl>     <dbl>
##  1 Abilene    2007  167.   19338549. 103650      864.      4.96
##  2 Abilene    2008  138.   16043361. 107133.     924       6.32
##  3 Abilene    2009  136.   16863146. 108367.     812.      6.12
##  4 Amarillo   2007  275    37922917. 117817.    1305.      4.55
##  5 Amarillo   2008  251.   36051667. 124000     1460.      5.48
##  6 Amarillo   2009  233.   33036863. 123283.    1477.      6.49
##  7 Arlington  2007  490.   72955011. 130283.    3030.      5.73
##  8 Arlington  2008  401.   59170193  129542.    2438.      5.58
##  9 Arlington  2009  356.   51298264. 128758.    1923.      5.29
## 10 Austin     2007 2337. 575913540  183650     9833.      3.98
```

9. Mean Estimation and Confidence Interval

```r
t.test(txhousing_clean$median, mu = 0)
```

```
##
##  One Sample t-test
##
## data:  txhousing_clean$median
## t = 225.8, df = 8601, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  117923.1 119988.5
## sample estimates:
## mean of x
##  118955.8
```

```r
lm(median ~ 1, data = txhousing_clean) %>% summary()
```

```
##
## Call:
## lm(formula = median ~ 1, data = txhousing_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -118956  -25156    1044   28944  185244
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 118955.8      526.8   225.8   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 48860 on 8601 degrees of freedom
```