Real Estate Pricing Data

Kady Barnes

05/01/2023

Real Estate Data

My husband is in the military, so we tend to move often. We are getting to the point where we will want to purchase a house, so I want to analyze housing data to see what type of relationships there are in housing prices and other factors like time on the market, time of year, and total listings. While we may not have a lot of choice in where we go, there might be a better time to buy!

```
library(tidyverse)
library(lubridate)
library(semTools)
library(ggplot2)
library(gridExtra)

Real_Estate_State <- read.csv(file = "Data/RDC_Inventory_Core_Metrics_State_History.csv",
    stringsAsFactors = TRUE)
```

General Cleanup

- Review Data Types
- Review NAs
- Review Variables

Determine if NAs should be removed

```
sum(is.na(Real_Estate_State))
```

```
## [1] 15463
```

```
total.na <- Real_Estate_State %>%
    group_by(Real_Estate_State$state) %>%
    summarize(sum(is.na(Real_Estate_State$median_listing_price_mm)))
total.na
```

```
## # A tibble: 52 x 2
##    `Real_Estate_State$state` sum(is.na(Real_Estate_State$median_listing_price_~1
##    <fct>                                                                  <int>
##  1 alabama                                                                  613
```

```
##  2 alaska                                                                          613
##  3 arizona                                                                         613
##  4 arkansas                                                                        613
##  5 california                                                                      613
##  6 colorado                                                                        613
##  7 connecticut                                                                     613
##  8 delaware                                                                        613
##  9 district of columbia                                                            613
## 10 florida                                                                         613
## # ... with 42 more rows, and abbreviated variable name
## #   1: `sum(is.na(Real_Estate_State$median_listing_price_mm))`
```

```
Real_Estate_Cleaned <- na.omit(Real_Estate_State)
sum(is.na(Real_Estate_Cleaned))
```

```
## [1] 0
```

```
State.Count <- table(Real_Estate_State$state)
State.Count
```

```
##
##              alabama               alaska              arizona
##                   74                   74                   74
##             arkansas           california             colorado
##                   74                   74                   74
##          connecticut             delaware district of columbia
##                   74                   74                   74
##              florida              georgia               hawaii
##                   74                   74                   74
##                idaho             illinois              indiana
##                   74                   74                   74
##                 iowa               kansas             kentucky
##                   74                   74                   74
##            louisiana                maine     marshall islands
##                   74                   74                    1
##             maryland        massachusetts             michigan
##                   74                   74                   74
##            minnesota          mississippi             missouri
##                   74                   74                   74
##              montana             nebraska               nevada
##                   74                   74                   74
##        new hampshire           new jersey           new mexico
##                   74                   74                   74
##             new york       north carolina         north dakota
##                   74                   74                   74
##                 ohio             oklahoma               oregon
##                   74                   74                   74
##         pennsylvania         rhode island       south carolina
##                   74                   74                   74
##         south dakota            tennessee                texas
##                   74                   74                   74
##                 utah              vermont             virginia
##                   74                   74                   74
```

```
##          washington       west virginia           wisconsin
##                  74                  74                  74
##             wyoming
##                  74
```

```r
State.Count.Cleaned <- table(Real_Estate_Cleaned$state)
State.Count.Cleaned
```

```
##
##             alabama              alaska             arizona
##                  62                  30                  62
##            arkansas          california            colorado
##                  62                  62                  62
##         connecticut            delaware district of columbia
##                  62                  62                  62
##             florida             georgia              hawaii
##                  62                  62                  62
##               idaho            illinois             indiana
##                  62                  62                  62
##                iowa              kansas            kentucky
##                  62                  62                  62
##           louisiana               maine    marshall islands
##                  62                  62                   0
##            maryland       massachusetts            michigan
##                  62                  62                  62
##           minnesota         mississippi            missouri
##                  62                  62                  62
##             montana            nebraska              nevada
##                  62                  62                  62
##       new hampshire          new jersey          new mexico
##                  62                  62                  62
##            new york      north carolina        north dakota
##                  62                  62                  62
##                ohio            oklahoma              oregon
##                  62                  62                  62
##        pennsylvania        rhode island      south carolina
##                  62                  62                  62
##        south dakota           tennessee               texas
##                  62                  62                  62
##                utah             vermont            virginia
##                  62                  62                  62
##          washington       west virginia           wisconsin
##                  62                  62                  62
##             wyoming
##                  62
```

Results:

- N/A's: 15,463; many of these are from the same variables to include the mm and yy changes- they do not appear to be random. Many of the categories have 613 NA's, which indicates they could potentially have been left out deliberately. For example, the first column with NA's is 'median_listing_price_mm.' Each state has 613 values missing.

- We may consider using summary statistics and fill missing values with the average or median for that specific state. However, as the missing data is relatively consistent, removing the NA's will not adversely affect one state more so than the other. Therefore, we will remove the NA's from the data.
- This removes a consistent amount from each state, confirming that removing the NA's will not throw off data for any one specific state. The exception is the Marshall Islands that had one data, but now has zero.

Review Data Types

```
# lapply(Real_Estate_Cleaned, class)

dates <- ym(Real_Estate_Cleaned$month_date_yyyymm)
str(dates)
```

```
##  Date[1:3130], format: "2022-08-01" "2022-07-01" "2022-06-01" "2022-05-01" "2022-04-01" ...
```

```
class(dates)
```

```
## [1] "Date"
```

Results:

- The only category that needs to be updated is the date formatting.

Review Variables

- All variables appear normal with the exception of: Quality Flag. "Triggered ("1") when data values are outside of their typical range. While rare, these figures should be reviewed before reporting."

```
# identify how many rows these affect:

length(which(Real_Estate_Cleaned$quality_flag == 0))
```

```
## [1] 3067
```

```
length(which(Real_Estate_Cleaned$quality_flag == 1))
```

```
## [1] 63
```

```
# There are 63 values that are listed as 1- meaning these are outside
# their typical range.  The data library does not state specifically
# what variable triggered the potential outlier.

grouped_quality_flag <- Real_Estate_Cleaned %>%
    group_by(state) %>%
    summarize(quality_flag = length(which(quality_flag == 1))) %>%
    arrange(desc(quality_flag))
grouped_quality_flag
```

```
## # A tibble: 51 x 2
##    state                quality_flag
##    <fct>                       <int>
##  1 district of columbia           11
##  2 utah                            6
##  3 arizona                         4
##  4 idaho                           4
##  5 michigan                        4
##  6 washington                      4
##  7 maine                           3
##  8 massachusetts                   3
##  9 new jersey                      3
## 10 pennsylvania                    3
## # ... with 41 more rows
```

Results:

- The area with the highest potential outliers ('1') is Washington, D.C. with 11. Deleting these rows will reduce D.C. from 62 to 51, removing approximately 18% of its data. Additionally, 6 rows from Utah will be removed, accounting for approximately 10% of its data. At this time, I will not remove the quality flags that equal 1. I will monitor the results to see if there are any trends in these two states specifically that may give us any insights.

Grouping Data to Begin Identifying Trends

- Are there any trends that can be seen with the data before I graph them?

States and Dates

```
# Grouping by date and state

Date.df.yr <- data.frame(date = c(format(dates, "%y")), average.price = c(Real_Estate_Cleaned$average_li

Date.df.month <- data.frame(date = c(format(dates, "%b")), average.price = c(Real_Estate_Cleaned$average

Date.df.yr %>%
    group_by(month = lubridate::floor_date(dates, "%y")) %>%
    summarize(average_price = mean(x = average.price)[1])
```

```
## # A tibble: 6 x 2
##   month      average_price
##   <date>             <dbl>
## 1 2017-01-01       451630.
## 2 2018-01-01       466341.
## 3 2019-01-01       483197.
## 4 2020-01-01       537360.
## 5 2021-01-01       610887.
## 6 2022-01-01       660138.
```

```r
Date.df.month %>%
    group_by(month = lubridate::floor_date(dates, "%b")) %>%
    summarize(average_price = mean(x = average.price)[1])
```

```
## # A tibble: 31 x 2
##    month      average_price
##    <date>             <dbl>
##  1 2017-07-01       453538.
##  2 2017-09-01       450860.
##  3 2017-11-01       450511.
##  4 2018-01-01       461389.
##  5 2018-03-01       473226.
##  6 2018-05-01       476685.
##  7 2018-07-01       468266.
##  8 2018-09-01       460742.
##  9 2018-11-01       457827.
## 10 2019-01-01       465271.
## # ... with 21 more rows
```

Results:

- Sales have continued to increase steadily over the past several years; I thought there would be a dip in 2020, but there wasn't. There also does not seem to be large differences in average price and the month.

```r
# Divided the states into 4 regions (determined by the Census Bureau)
# and divided the months into 4 seasons to allow a better visual
# picture.

Real_Estate_Cleaned_Recode <- Real_Estate_Cleaned %>%
    mutate(state = recode(.x = state, connecticut = "Northeast", maine = "Northeast",
        massachusetts = "Northeast", `new hampshire` = "Northeast", `rhode island` = "Northeast",
        vermont = "Northeast", `new jersey` = "Northeast", `new york` = "Northeast",
        pennsylvania = "Northeast")) %>%
    mutate(state = recode(.x = state, illinois = "Midwest", indiana = "Midwest",
        michigan = "Midwest", ohio = "Midwest", wisconsin = "Midwest",
        iowa = "Midwest", kansas = "Midwest", minnesota = "Midwest", missouri = "Midwest",
        nebraska = "Midwest", `north dakota` = "Midwest", `south dakota` = "Midwest")) %>%
    mutate(state = recode(.x = state, delaware = "South", florida = "South",
        georgia = "South", maryland = "South", `north carolina` = "South",
        `south carolina` = "South", virginia = "South", `district of columbia` = "South",
        `west virginia` = "South", alabama = "South", kentucky = "South",
        mississippi = "South", tennessee = "South", arkansas = "South",
        louisiana = "South", oklahoma = "South", texas = "South")) %>%
    mutate(state = recode(.x = state, arizona = "West", colorado = "West",
        idaho = "West", montana = "West", nevada = "West", `new mexico` = "West",
        utah = "West", wyoming = "West", alaska = "West", california = "West",
        hawaii = "West", oregon = "West", washington = "West")) %>%
    mutate(state = recode(.x = state, `marshall islands` = "Other")) %>%
```

```r
    mutate(dates = as.factor(dates)) %>%
    mutate(dates = recode_factor(.x = dates, `2022-12-01` = "Winter", `2021-12-01` = "Winter",
        `2020-12-01` = "Winter", `2019-12-01` = "Winter", `2018-12-01` = "Winter",
        `2017-12-01` = "Winter", `2016-12-01` = "Winter", `2022-01-01` = "Winter",
        `2021-01-01` = "Winter", `2020-01-01` = "Winter", `2019-01-01` = "Winter",
        `2018-01-01` = "Winter", `2017-01-01` = "Winter", `2016-01-01` = "Winter",
        `2022-02-01` = "Winter", `2021-02-01` = "Winter", `2020-02-01` = "Winter",
        `2019-02-01` = "Winter", `2018-02-01` = "Winter", `2017-02-01` = "Winter",
        `2016-02-01` = "Winter", `2022-03-01` = "Spring", `2021-03-01` = "Spring",
        `2020-03-01` = "Spring", `2019-03-01` = "Spring", `2018-03-01` = "Spring",
        `2017-03-01` = "Spring", `2016-03-01` = "Spring", `2022-04-01` = "Spring",
        `2021-04-01` = "Spring", `2020-04-01` = "Spring", `2019-04-01` = "Spring",
        `2018-04-01` = "Spring", `2017-04-01` = "Spring", `2016-04-01` = "Spring",
        `2022-05-01` = "Spring", `2021-05-01` = "Spring", `2020-05-01` = "Spring",
        `2019-05-01` = "Spring", `2018-05-01` = "Spring", `2017-05-01` = "Spring",
        `2016-05-01` = "Spring", `2022-06-01` = "Summer", `2021-06-01` = "Summer",
        `2020-06-01` = "Summer", `2019-06-01` = "Summer", `2018-06-01` = "Summer",
        `2017-06-01` = "Summer", `2016-06-01` = "Summer", `2022-07-01` = "Summer",
        `2021-07-01` = "Summer", `2020-07-01` = "Summer", `2019-07-01` = "Summer",
        `2018-07-01` = "Summer", `2017-07-01` = "Summer", `2016-07-01` = "Summer",
        `2022-08-01` = "Summer", `2021-08-01` = "Summer", `2020-08-01` = "Summer",
        `2019-08-01` = "Summer", `2018-08-01` = "Summer", `2017-08-01` = "Summer",
        `2016-08-01` = "Summer", `2022-09-01` = "Fall", `2021-09-01` = "Fall",
        `2020-09-01` = "Fall", `2019-09-01` = "Fall", `2018-09-01` = "Fall",
        `2017-09-01` = "Fall", `2016-09-01` = "Fall", `2022-10-01` = "Fall",
        `2021-10-01` = "Fall", `2020-10-01` = "Fall", `2019-10-01` = "Fall",
        `2018-10-01` = "Fall", `2017-10-01` = "Fall", `2016-10-01` = "Fall",
        `2022-11-01` = "Fall", `2021-11-01` = "Fall", `2020-11-01` = "Fall",
        `2019-11-01` = "Fall", `2018-11-01` = "Fall", `2017-11-01` = "Fall",
        `2016-11-01` = "Fall"))


summary(Real_Estate_Cleaned_Recode$dates)
```

Divide the states into 4 regions (determined by the Census Bureau) and the months into four seasons for a better picture when graphed and regressions performed.

```
## Winter Spring Summer   Fall
##    759    756    856    759
```

```r
summary(Real_Estate_Cleaned_Recode$state)
```

```
##     South      West Northeast   Midwest     Other
##      1054       774       558       744         0
```

Results:

- We see that the Summer has the highest entries and the remaining seasons have relatively the lowest entries. This does not really tell us much yet without comparing it do a different variable.
- This also does not tell us much without comparing it to another variable, but the South has highest value here.

Average Listing Price

```
sort(tapply(Real_Estate_Cleaned$average_listing_price, Real_Estate_Cleaned$state,
    mean), decreasing = TRUE)
```

```
##               hawaii           california             new york
##            1396856.6            1237778.0            1148788.0
##             colorado        massachusetts district of columbia
##            1013154.8             968600.4             960824.2
##                 utah          connecticut              montana
##             864572.6             864351.3             765997.4
##              florida           washington                idaho
##             741469.1             692108.1             683032.3
##         rhode island               nevada              wyoming
##             669755.4             664886.9             644220.5
##               oregon              arizona           new jersey
##             615238.2             611128.4             599652.2
##        new hampshire             virginia              vermont
##             502368.9             489351.2             473305.9
##             maryland       north carolina       south carolina
##             471650.2             452354.9             447518.2
##                texas             delaware              georgia
##             446521.9             445312.2             441811.4
##            tennessee                maine           new mexico
##             437615.5             423963.9             417601.2
##               alaska            minnesota             illinois
##             404030.3             398656.5             391236.9
##            wisconsin         pennsylvania             michigan
##             351383.8             347753.2             332483.3
##         south dakota            louisiana              alabama
##             330621.3             330160.2             329974.5
##             nebraska             missouri             oklahoma
##             310483.9             304021.5             302626.8
##             kentucky         north dakota               kansas
##             302526.6             288690.0             288236.5
##             arkansas              indiana          mississippi
##             284992.3             283228.5             282086.1
##                 ohio                 iowa        west virginia
##             276877.0             259470.8             247042.8
```

```
sort(tapply(Real_Estate_Cleaned$average_listing_price, Real_Estate_Cleaned$state,
    median), decreasing = TRUE)
```

```
##               hawaii           california             new york
##            1339815.0            1128075.5            1103088.0
## district of columbia             colorado        massachusetts
##             956219.5             941423.0             865405.0
##          connecticut                 utah              florida
##             752397.5             746167.5             671833.0
##           washington              montana              wyoming
##             664034.0             653164.5             631421.5
##               nevada         rhode island                idaho
```

```
##            625131.0            616304.5            593955.5
##             arizona          new jersey              oregon
##            582987.5            575697.0            567380.0
##            maryland            virginia       new hampshire
##            462536.0            461840.0            459653.5
##            delaware      south carolina      north carolina
##            430677.0            426735.5            425855.5
##               texas             vermont              alaska
##            421884.5            419614.5            406192.5
##             georgia           tennessee          new mexico
##            404593.0            397662.0            389037.5
##           minnesota               maine            illinois
##            387530.5            387246.0            381605.5
##           wisconsin        pennsylvania            michigan
##            342729.0            325992.0            322438.5
##             alabama        south dakota           louisiana
##            312754.0            309751.0            306143.5
##            nebraska            kentucky            oklahoma
##            293863.0            288023.0            286874.0
##            missouri              kansas        north dakota
##            286823.0            285210.5            281354.5
##             indiana                ohio            arkansas
##            276945.5            273427.5            265970.5
##         mississippi                iowa       west virginia
##            264499.5            254540.5            233084.0
```

```r
# Grouped by region:

State.Prices <- Real_Estate_Cleaned_Recode %>%
    group_by(state) %>%
    summarize(average_price = mean(x = average_listing_price)[1])
State.Prices
```

```
## # A tibble: 4 x 2
##   state     average_price
##   <fct>             <dbl>
## 1 South           436108.
## 2 West            785179.
## 3 Northeast       666504.
## 4 Midwest         317949.
```

```r
# Breaking down by season instead of specific months:

Season.Prices <- Real_Estate_Cleaned_Recode %>%
    group_by(dates) %>%
    summarize(average_price = mean(x = average_listing_price)[1])
Season.Prices
```

```
## # A tibble: 4 x 2
##   dates  average_price
##   <fct>          <dbl>
## 1 Winter       530091.
## 2 Spring       553256.
```

```
## 3 Summer       543722.
## 4 Fall         513602.
```

Results:

- Spring has the highest average price, with the lowest being Fall. This is interesting as I would have expected that Winter would be the lowest due to the cold. Although warmer climate regions like the West and South could be a contributing factor (selling a house in the fall or winter in these regions when it is not so hot). Additionally, the West has the highest average price, while the Midwest has the least.

Median Days on Market

```
# Grouped by state:

sort(tapply(Real_Estate_Cleaned$median_days_on_market, Real_Estate_Cleaned$state,
    mean), decreasing = TRUE)
```

```
##              vermont                maine              montana
##            110.90323             91.62903             88.59677
##        west virginia          mississippi              wyoming
##             87.20968             82.09677             80.87097
##             delaware            louisiana               hawaii
##             79.85484             75.32258             74.95161
##           new mexico             new york             arkansas
##             74.46774             74.27419             73.20968
##         north dakota              alabama       north carolina
##             73.09677             71.01613             69.69355
##              florida         pennsylvania       south carolina
##             68.58065             68.48387             68.29032
##            wisconsin               alaska             missouri
##             67.30645             64.83333             64.70968
##                 iowa          connecticut               kansas
##             64.53226             63.85484             63.72581
##             kentucky             oklahoma        new hampshire
##             63.64516             60.50000             60.43548
##             michigan             virginia                texas
##             57.77419             57.43548             56.22581
##             colorado           new jersey              indiana
##             55.30645             54.91935             54.83871
##         south dakota             illinois              georgia
##             54.70968             54.64516             54.50000
##            minnesota               oregon                 ohio
##             54.43548             54.27419             54.03226
##            tennessee             maryland                idaho
##             52.74194             52.66129             52.41935
##        massachusetts             nebraska         rhode island
##             51.59677             51.41935             50.98387
##              arizona                 utah               nevada
##             49.51613             49.32258             44.54839
##           washington           california district of columbia
##             43.38710             42.35484             39.20968
```

```r
sort(tapply(Real_Estate_Cleaned$median_days_on_market, Real_Estate_Cleaned$state,
    median), decreasing = TRUE)
```

```
##              vermont        west virginia          mississippi
##                108.0                 90.0                 88.0
##                maine              montana             delaware
##                 86.5                 86.5                 84.0
##               hawaii             arkansas            louisiana
##                 80.0                 78.0                 78.0
##              wyoming              florida           new mexico
##                 78.0                 75.0                 75.0
##             new york              alabama       south carolina
##                 74.0                 73.0                 72.5
##       north carolina         north dakota         pennsylvania
##                 71.0                 71.0                 69.0
##             kentucky            wisconsin               alaska
##                 66.0                 65.5                 65.0
##                 iowa             missouri               kansas
##                 65.0                 65.0                 64.0
##          connecticut             oklahoma        new hampshire
##                 61.0                 61.0                 58.5
##                texas              georgia             virginia
##                 57.0                 56.0                 56.0
##             michigan           new jersey              indiana
##                 55.5                 55.5                 55.0
##                 ohio            tennessee        massachusetts
##                 55.0                 54.5                 54.0
##              arizona             illinois             maryland
##                 53.5                 53.5                 52.5
##         south dakota             colorado               oregon
##                 52.5                 52.0                 51.5
##         rhode island             nebraska            minnesota
##                 51.5                 51.0                 50.5
##                 utah                idaho               nevada
##                 50.5                 49.0                 44.0
##           california           washington district of columbia
##                 42.0                 39.0                 37.0
```

```r
# Grouped by region:

State.Days <- Real_Estate_Cleaned_Recode %>%
    group_by(state) %>%
    summarize(median_days = mean(x = median_days_on_market)[1])
State.Days
```

```
## # A tibble: 4 x 2
##   state     median_days
##   <fct>           <dbl>
## 1 South            65.4
## 2 West             59.4
## 3 Northeast        69.7
## 4 Midwest          59.6
```

```r
# Grouped by season:

Season.Days <- Real_Estate_Cleaned_Recode %>%
    group_by(dates) %>%
    summarize(median_days = mean(x = median_days_on_market)[1])
Season.Days
```

```
## # A tibble: 4 x 2
##   dates  median_days
##   <fct>        <dbl>
## 1 Winter        80.0
## 2 Spring        56.5
## 3 Summer        53.7
## 4 Fall          64.3
```

Results:

- #The median days on the market are relatively similar all around, with the South leading, and winter has the highest median days on the market.

New Listing Count

```r
# Grouped by state:

sort(tapply(Real_Estate_Cleaned$new_listing_count, Real_Estate_Cleaned$state,
    mean), decreasing = TRUE)
```

```
##           florida             texas        california
##        43958.0645        38706.0645        38580.4516
##           georgia          illinois          new york
##        25170.1290        20013.9355        17064.8387
##    north carolina              ohio        new jersey
##        14916.9677        14511.2903        14211.7419
##      pennsylvania          michigan          virginia
##        13924.0000        13379.3548        12687.3548
##           arizona          colorado         tennessee
##        11797.5484        10888.5161        10744.8387
##        washington    south carolina          missouri
##         9804.5161         8929.1613         8648.3226
##         minnesota          maryland           indiana
##         8292.7742         8023.2903         7802.3871
##     massachusetts         wisconsin            oregon
##         7718.9677         7105.1613         6417.7419
##           alabama              utah          kentucky
##         6332.9677         5248.3871         5246.4516
##          oklahoma       connecticut         louisiana
##         4967.0968         4904.0000         4847.8710
##            nevada              iowa          arkansas
##         4840.3871         4369.2903         3914.3226
##            kansas             idaho       mississippi
```

```
##            3556.9032            3245.2258            2612.3871
##           new mexico             nebraska        new hampshire
##            2516.5161            2411.8710            1931.7419
##                maine        west virginia               hawaii
##            1863.9355            1700.0645            1688.5806
##              montana         rhode island             delaware
##            1680.2581            1451.6774            1344.5161
##         south dakota district of columbia         north dakota
##            1075.0968             998.9677             891.0968
##              wyoming              vermont               alaska
##             870.1290             799.6774             648.5333
```

```r
sort(tapply(Real_Estate_Cleaned$new_listing_count, Real_Estate_Cleaned$state,
    median), decreasing = TRUE)
```

```
##              florida           california                texas
##                43824                39548                38174
##              georgia             illinois             new york
##                25306                20474                17582
##                 ohio       north carolina           new jersey
##                15380                15334                14988
##         pennsylvania             michigan             virginia
##                14514                13954                12910
##              arizona             colorado            tennessee
##                11832                11046                10882
##           washington       south carolina             missouri
##                10432                 9094                 9010
##            minnesota              indiana             maryland
##                 8864                 8174                 8170
##        massachusetts            wisconsin               oregon
##                 8068                 7482                 6552
##              alabama                 utah             kentucky
##                 6462                 5476                 5370
##          connecticut             oklahoma               nevada
##                 5120                 5008                 4916
##            louisiana                 iowa             arkansas
##                 4908                 4492                 3928
##               kansas                idaho          mississippi
##                 3708                 3216                 2646
##           new mexico             nebraska        new hampshire
##                 2552                 2536                 1946
##                maine        west virginia               hawaii
##                 1848                 1750                 1694
##              montana         rhode island             delaware
##                 1646                 1492                 1364
##         south dakota district of columbia         north dakota
##                 1144                 1008                  900
##              wyoming              vermont               alaska
##                  872                  778                  610
```

```r
# Grouped by region:
State.Listing <- Real_Estate_Cleaned_Recode %>%
    group_by(state) %>%
```

```r
    summarize(new_listing = mean(x = new_listing_count)[1])
State.Listing
```

```
## # A tibble: 4 x 2
##   state     new_listing
##   <fct>           <dbl>
## 1 South          11477.
## 2 West            7841.
## 3 Northeast       7097.
## 4 Midwest         7671.
```

```r
# Grouped by season:

Season.Listing <- Real_Estate_Cleaned_Recode %>%
    group_by(dates) %>%
    summarize(new_listing = mean(x = new_listing_count)[1])
Season.Listing
```

```
## # A tibble: 4 x 2
##   dates  new_listing
##   <fct>        <dbl>
## 1 Winter       6933.
## 2 Spring       9787.
## 3 Summer      10164.
## 4 Fall         8526.
```

Results:

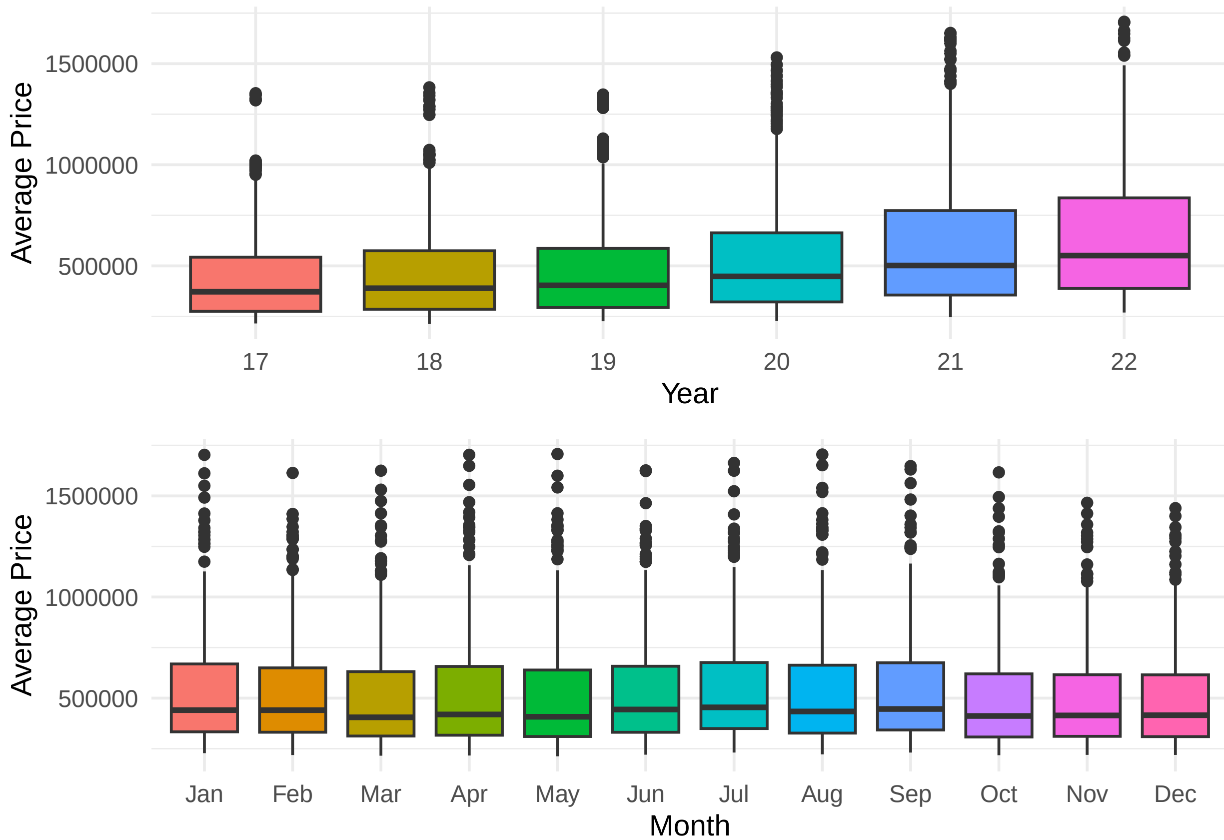
- The South has the most new listings, which makes adds more context as a contributing factor as to why from #3 that it has the longest median days on the market. The Summer has the highest total listings and the Winter has the least new total listings.
- Going back to the quality flag variable where D.C. stood out the most, we see that DC has the second lowest total listing count, but the sixth highest average listing price, and the lowest average days on the market. While it makes sense that a smaller area would have less houses available and likely increase demand, this can be a factor that would cause the quality flag to rise.

Graphing Relationships for Visual Representation

Date Versus Average Listing Price:

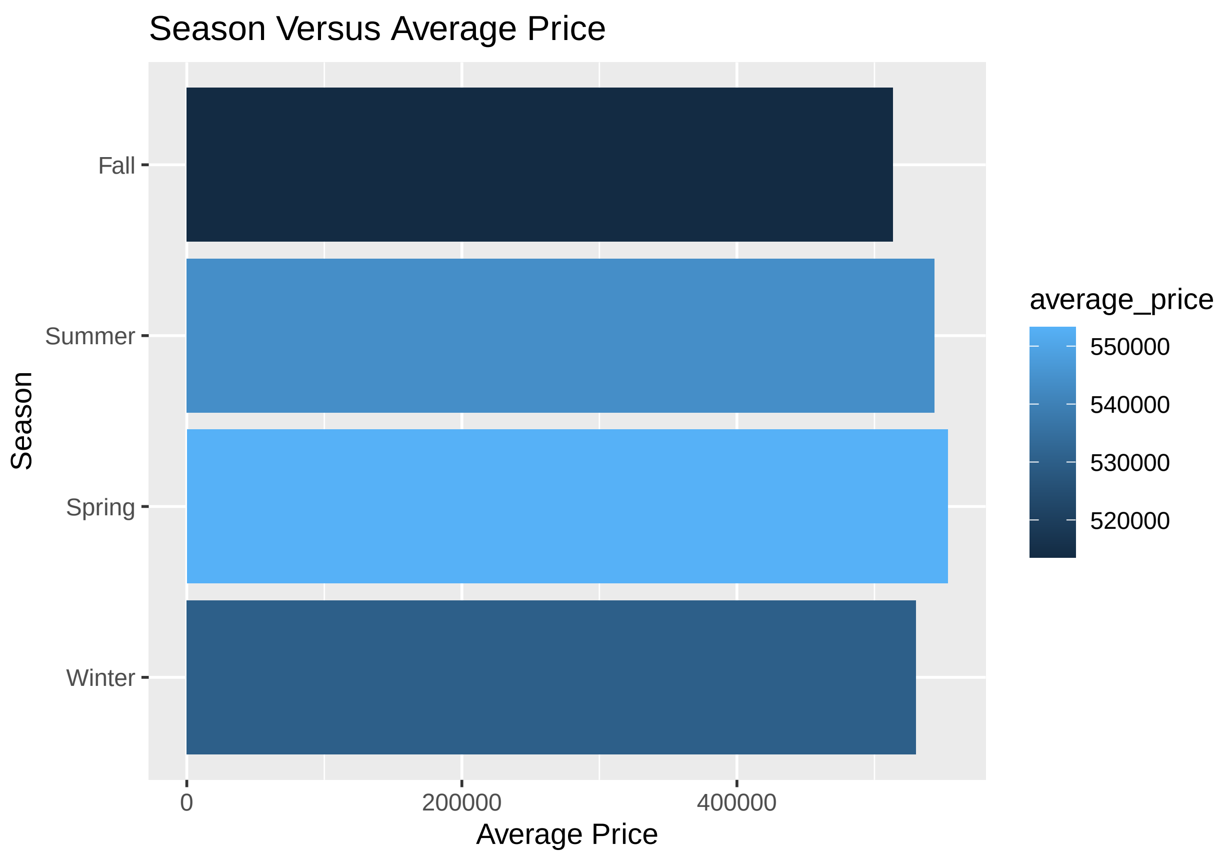
```r
Plot1 <- Date.df.yr %>%
    ggplot(aes(y = average.price, x = date, fill = date)) + geom_boxplot() +
    theme_minimal() + labs(x = "Year", y = "Average Price") + theme(legend.position = "none")


Plot2 <- Date.df.month %>%
    ggplot(aes(y = average.price, x = date, fill = date)) + geom_boxplot() +
    theme_minimal() + labs(x = "Month", y = "Average Price") + theme(legend.position = "none") +
```

```
    scale_x_discrete(labels = month.abb)

gridExtra::grid.arrange(Plot1, Plot2)
```



Season Versus Average Listing Price:

```
options(scipen = 6)

ggplot(Season.Prices, aes(dates, average_price, fill = average_price)) +
    geom_bar(stat = "identity", position = "dodge") + coord_flip() + labs(x = "Season",
    y = "Average Price") + ggtitle("Season Versus Average Price")
```



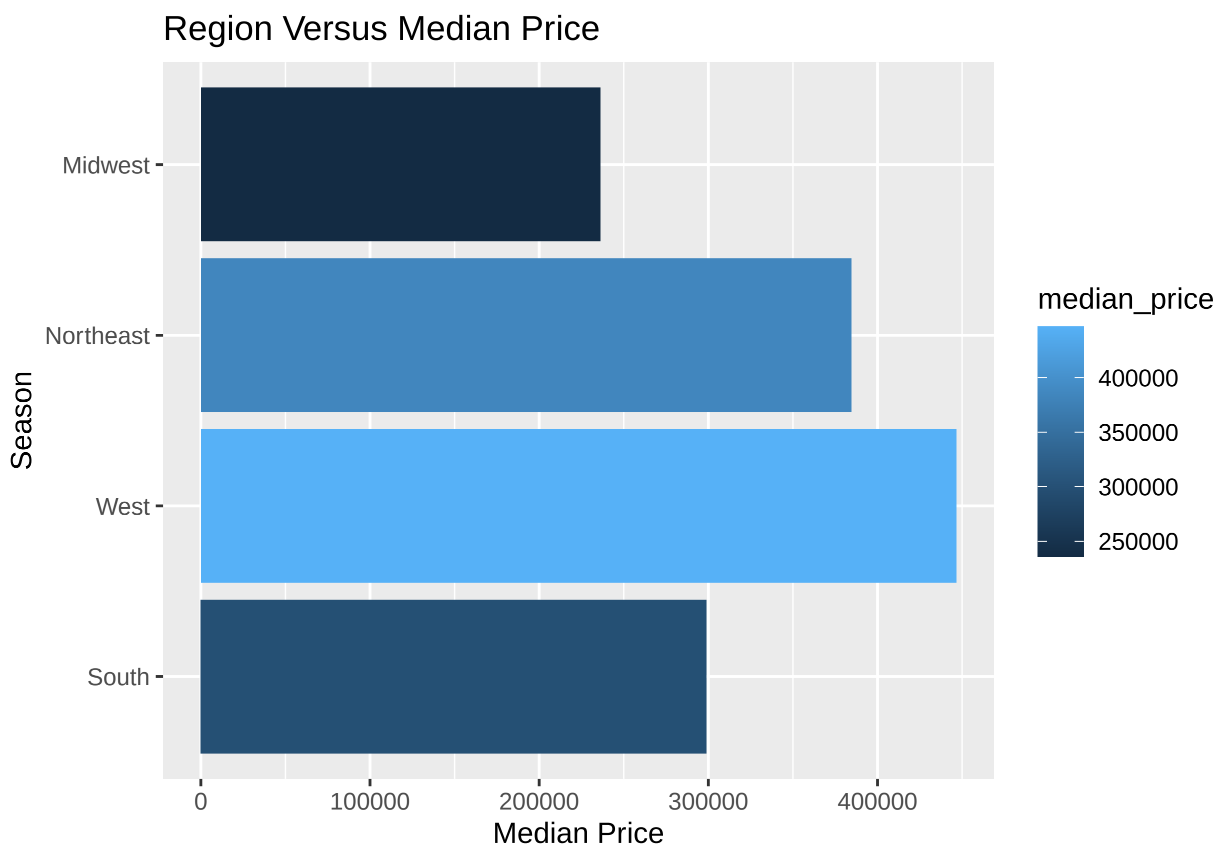
Results: - This shows us a general increase in the average price over the years, but (at least visually), not a large difference in average price and month and season.

Median List Price Versus Regions:

```
Region.Prices <- Real_Estate_Cleaned_Recode %>%
    group_by(state) %>%
    summarize(median_price = mean(x = median_listing_price)[1])
Region.Prices
```

```
## # A tibble: 4 x 2
##   state     median_price
##   <fct>            <dbl>
## 1 South          298856.
## 2 West           446489.
## 3 Northeast      384417.
## 4 Midwest        236142.
```

```
ggplot(Region.Prices, aes(state, median_price, fill = median_price)) +
    geom_bar(stat = "identity", position = "dodge") + coord_flip() + labs(x = "Season",
    y = "Median Price") + ggtitle("Region Versus Median Price")
```



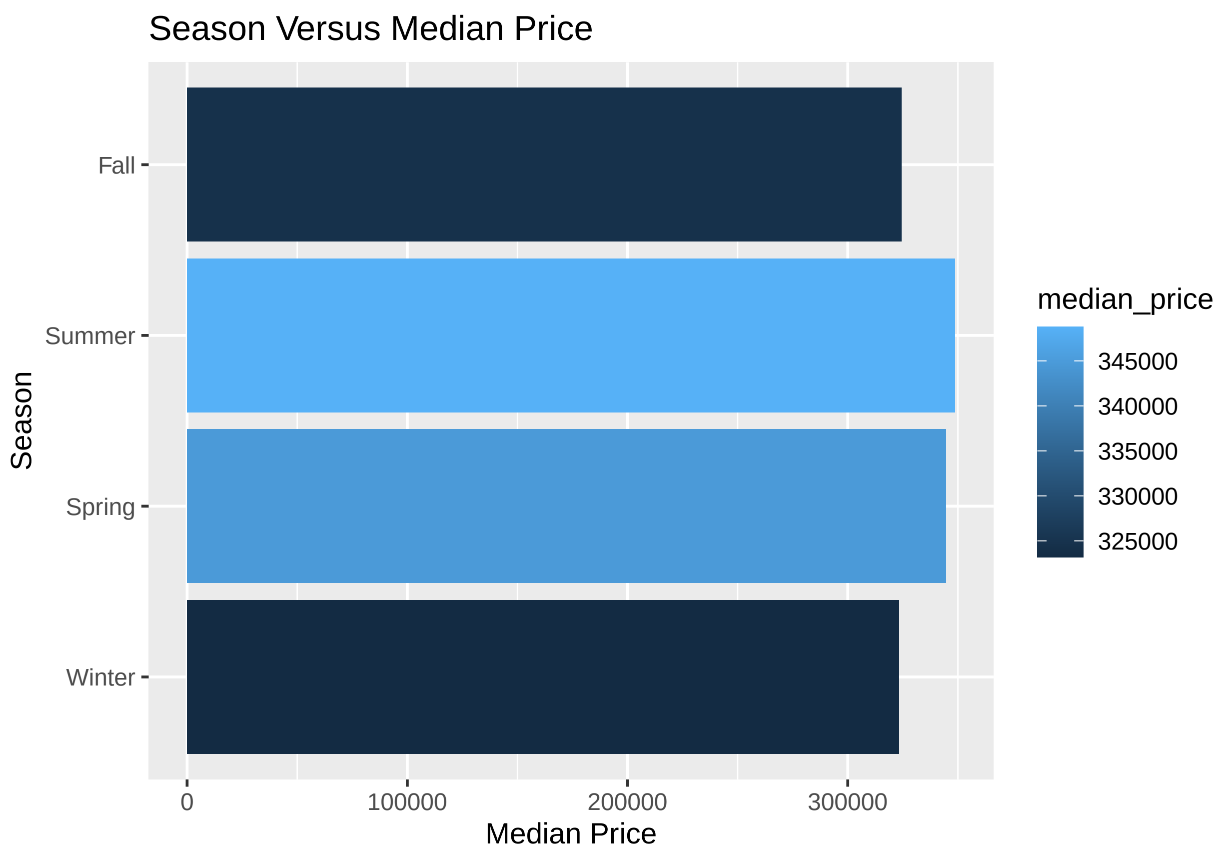
Results: - The West and Northeast have the highest average house prices with the Midwest coming in at the lowest average house prices.

Median List Price Versus Seasons:

```
Season.Median.Prices <- Real_Estate_Cleaned_Recode %>%
    group_by(dates) %>%
    summarize(median_price = mean(x = median_listing_price)[1])
Season.Median.Prices
```

```
## # A tibble: 4 x 2
##   dates  median_price
##   <fct>         <dbl>
## 1 Winter      323239.
## 2 Spring      344726.
## 3 Summer      348757.
## 4 Fall        324484.
```

```
ggplot(Season.Median.Prices, aes(dates, median_price, fill = median_price)) +
    geom_bar(stat = "identity", position = "dodge") + coord_flip() + labs(x = "Season",
    y = "Median Price") + ggtitle("Season Versus Median Price")
```

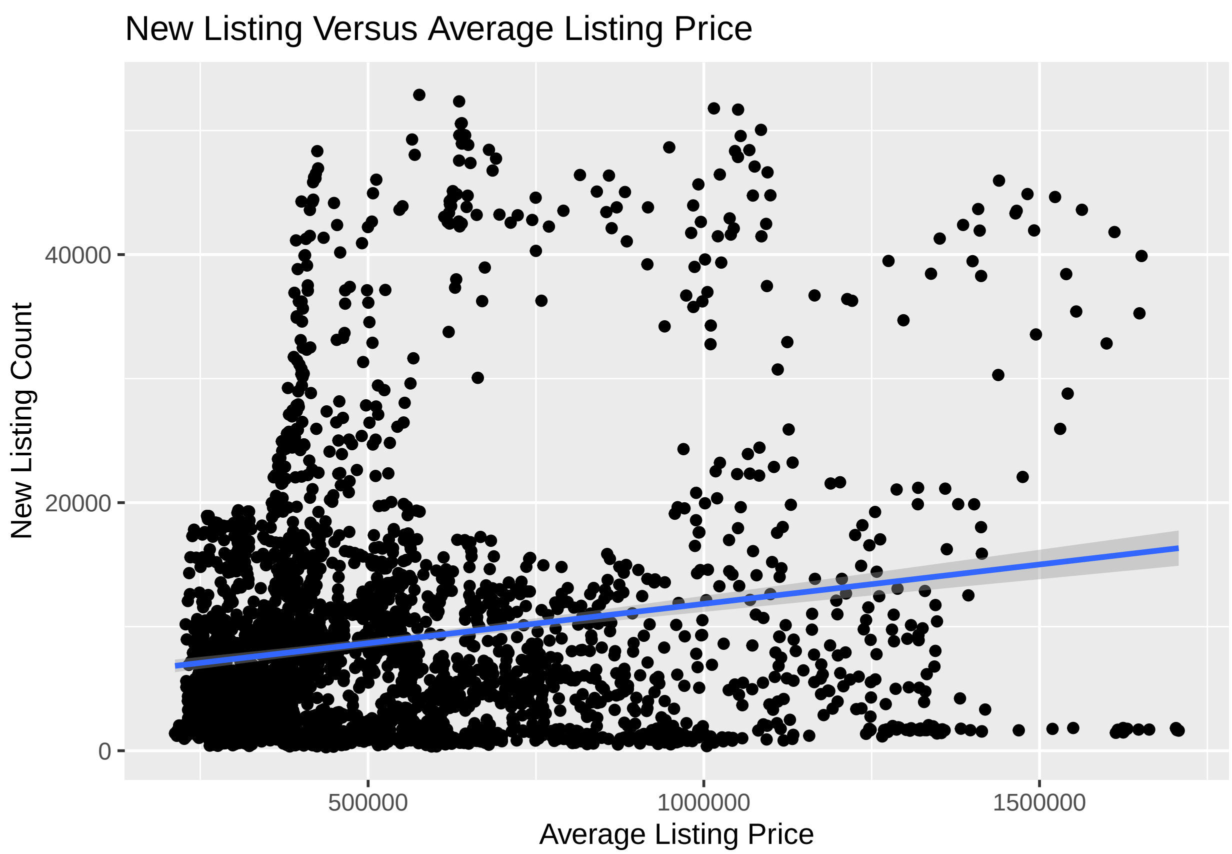


Results: - The West and Northeast have the highest average house prices with the Midwest coming in at the lowest average house prices.

Average Listing Price Versus New Listing Count:

```
AvgPrice.NewList <- Real_Estate_Cleaned %>%
    ggplot(aes(x = average_listing_price, y = new_listing_count)) + geom_point() +
    stat_smooth(method = "lm") + labs(x = "Average Listing Price", y = "New Listing Count") +
    ggtitle("New Listing Versus Average Listing Price")
AvgPrice.NewList
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Results: - This graph shows a positive relationship between the two variables; as the average listing price increases, the new listing count increases as well; however, there are many outliers that could affect this relationship when looking at it statistically.

Regression Testing

I will test to see if there are any statistically significant variables with the Median List Price and the seasons and regions, as I believe these may have a large impact on the list price.

Median List Price with Regions and Seasons

I will test the differences in the the mean of the median listing price across seasons and region. Hypothesis 1:

- H0: The mean of the median list price is the same across all regions
- Ha: The mean of the median list price is not the same across all regions

Hypothesis 2:

- H0: The mean of the median list price is the same across all seasons
- Ha: The mean of the median list price is not the same across all seasons

Let Alpha = .05

```
Price2Way <- aov(median_listing_price ~ state + dates, data = Real_Estate_Cleaned_Recode)
summary(Price2Way)
```

```
##                Df         Sum Sq       Mean Sq F value       Pr(>F)
## state           3 19629748247865 6543249415955  594.21      < 2e-16 ***
## dates           3   439161824910  146387274970   13.29 0.0000000127 ***
## Residuals    3123 34389391238246   11011652654
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# P-Value for the four regions is < 2e-16

# P-Value for the four seasons is 1.27e-08

# The ANOVA results indicate there is a difference in both the mean
# of the median listing price compared to seasons and regions.
# Therefore, we reject the null hypotheses on both. I will conduct
# post-hoc tests to review further.

# Bonferroni Test for Hypothesis 1 (Regions):

pairwise.t.test(Real_Estate_Cleaned_Recode$median_listing_price, Real_Estate_Cleaned_Recode$state,
    p.adj = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Real_Estate_Cleaned_Recode$median_listing_price and Real_Estate_Cleaned_Recode$state
##
##           South  West   Northeast
## West      <2e-16 -      -
## Northeast <2e-16 <2e-16 -
## Midwest   <2e-16 <2e-16 <2e-16
##
## P value adjustment method: bonferroni
```

```
# The Bonferroni test shows us there is a statistically significant
# difference between the means of all regions and the median listing
# price.

# Tukey Test For Hypothesis 1 (Regions):

State1Way <- aov(median_listing_price ~ state, data = Real_Estate_Cleaned_Recode)

TukeyHSD(State1Way)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = median_listing_price ~ state, data = Real_Estate_Cleaned_Recode)
```

```
##
## $state
##                          diff        lwr        upr p adj
## West-South         147633.64  134790.43  160476.85     0
## Northeast-South     85561.45   71357.10   99765.80     0
## Midwest-South      -62713.28  -75704.94  -49721.63     0
## Northeast-West     -62072.19  -77139.68  -47004.71     0
## Midwest-West      -210346.93 -224277.06 -196416.79     0
## Midwest-Northeast -148274.73 -163468.95 -133080.52     0
```

```
# The Tukey test shows us that the West has the highest mean due its
# diff results with the other regions.  The second highest mean is
# Northeast, followed by the South, with the Midwest having the
# lowest mean.

# Bonferroni Test for Hypothesis 2 (Dates):

pairwise.t.test(Real_Estate_Cleaned_Recode$median_listing_price, Real_Estate_Cleaned_Recode$dates,
    p.adj = "bonf")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  Real_Estate_Cleaned_Recode$median_listing_price and Real_Estate_Cleaned_Recode$dates
##
##        Winter  Spring  Summer
## Spring 0.00891 -       -
## Summer 0.00061 1.00000 -
## Fall   1.00000 0.01653 0.00130
##
## P value adjustment method: bonferroni
```

```
# The Bonferroni test shows us a statistically significant difference
# in mean between Spring and Winter, Summer and Winter, Fall and
# Spring, and Fall and Summer.  This is interesting because this is
# quite a bit different than the statistically significant
# differences in the average list price.

Date1Way <- aov(median_listing_price ~ dates, data = Real_Estate_Cleaned_Recode)

TukeyHSD(Date1Way)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = median_listing_price ~ dates, data = Real_Estate_Cleaned_Recode)
##
## $dates
##                     diff        lwr       upr     p adj
## Spring-Winter  21487.178   4121.951 38852.406 0.0080850
## Summer-Winter  25518.309   8668.909 42367.709 0.0005852
## Fall-Winter     1245.209 -16102.816 18593.235 0.9977723
## Summer-Spring   4031.131 -12835.979 20898.242 0.9275568
```

```
## Fall-Spring   -20241.969 -37607.196 -2876.741 0.0146299
## Fall-Summer   -24273.100 -41122.500 -7423.700 0.0012403
```

```
# The Tukey Test shows us a difference in means with the median list
# price that we didn't see with the average list price test
# previously done.  This shows us that statistically significant
# difference in means is: Spring-Winter (with Spring being larger),
# Summer-Winter (With Summer being larger), Fall-Spring (with Spring
# being larger), and Fall-Summer, (with Summer being larger).
# Therefore, Spring has the largest average median listing price,
# with the Summer being the second largest.
```

Result: There were more statistically significant differences in means between the seasons than I thought. I initially only thought that the Summer would have statistically significant means due to the results of the average listing price tests, but we also had several other seasonal differences. Additionally, in regards to the regions, there was a statistically significant difference in all the regions, with the West being the highest, where I was correct in my prediction.

Correlation Testing

After completing the multiple regression, I wanted to see through a correlation test, how some of the numerical variables may affect the average listing price.

```
cor.test(Real_Estate_Cleaned_Recode$active_listing_count, Real_Estate_Cleaned_Recode$average_listing_pri
```

```
##
##  Pearson's product-moment correlation
##
## data:  Real_Estate_Cleaned_Recode$active_listing_count and Real_Estate_Cleaned_Recode$average_listing
## t = 3.6603, df = 3128, p-value = 0.0002561
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.0303399 0.1001121
## sample estimates:
##        cor
## 0.06530582
```

```
cor.test(Real_Estate_Cleaned_Recode$new_listing_count, Real_Estate_Cleaned_Recode$average_listing_price)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Real_Estate_Cleaned_Recode$new_listing_count and Real_Estate_Cleaned_Recode$average_listing_pr
## t = 10.567, df = 3128, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1516038 0.2192623
## sample estimates:
##       cor
## 0.1856531
```

```r
cor.test(Real_Estate_Cleaned_Recode$pending_ratio, Real_Estate_Cleaned_Recode$average_listing_price)
```

```
##
##  Pearson's product-moment correlation
##
## data:  Real_Estate_Cleaned_Recode$pending_ratio and Real_Estate_Cleaned_Recode$average_listing_price
## t = 12.121, df = 3128, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.1780953 0.2450260
## sample estimates:
##      cor
## 0.211809
```

```r
cor.test(Real_Estate_Cleaned_Recode$median_days_on_market, Real_Estate_Cleaned_Recode$average_listing_pr
```

```
##
##  Pearson's product-moment correlation
##
## data:  Real_Estate_Cleaned_Recode$median_days_on_market and Real_Estate_Cleaned_Recode$average_listin
## t = -15.661, df = 3128, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.3018322 -0.2368506
## sample estimates:
##        cor
## -0.2696483
```

```r
# The median days on the market seem to be most correlated with the
# average listing price.
```

Final Results

- The date and location seemed to be large predictors of price. Regarding dates, the confounding variables can be the variability in the climate. A listing in Florida in the Winter will likely be a lot different than a listing in the Midwest at the time same time due to extremely cold temperatures. Additionally, regarding regions, a factor that should be considered is the population density. More people located in an area equates to higher demand for housing, which increases the the housing price. An example of this is Washington D.C.; there seemed several outliers that flagged the quality flag indicator, but this is a small area that is densely populated; with this comes a higher demand for houses, thus allowing sellers to list homes for higher prices.
- Another factor that could affect the results is an expensive area that can affect the whole state. As an example, in New York, the most populous city is New York City (also the most populated city in the US). New York state's average listing price is the third highest in the country, yet travel upstate, and the housing prices will be less due to less demand and less job opportunities.

What does this mean for our potential housing purchase at our next duty station?

- The most likely moves will either be to California, Florida, Virginia, or Maryland. These are different geographic areas, and from the data, I know that California will be the highest price among the four states. Regardless of where we buy, the from both the visual graphs and the regression testing is that purchasing houses in the Spring will be the highest. While we usually move around the Spring and Summer, if we can hold out to a later season, we may find a house slightly cheaper in the Fall or Winter. Although, it is important to note that the only statistically significant price difference in seasons was between the Spring and Fall, so the Fall may be the way to go when looking to purchase a house.