Housing Sales in Texas
Exploratory Data Analysis
John Little

I. INTRODUCTION

The data set is built into R and contains a monthly summary of housing sales data in various Texas cities between 2000 and 2016. It includes data on the location, date, number of sales, selling price, number of listings, and how much inventory is available (in months of selling time). This dataset can be used to analyze market trends in the Texas housing market. The variables are as follows:

- 'city': name of the multiple listing service (MLS) area. The MLS area aligns closely to cities but is used by realtors to refer to the surrounding areas as well.
- 'year': year that the data was recorded
- 'month': month that the data was recorded. Months 1-12 represent January-December, respectively.
- 'sales': number of sales
- 'volume': total value of sales
- 'median': median sale price
- 'listings': total number of active listings
- 'inventory': amount of time it would take to sell all current listings at current rate

II. DATASET DESCRIPTION

The dataset contains 8602 rows and 9 columns with various data types. There are 2 'character' variables, 1 integer variable, and 6 numerical variables as shown in Figure 1.

Figure 1: Variable Names and Data Types

```
> ##dimensions of dataset
> dim(data)
[1] 8602    9

> ##variable names
> names(data)
[1] "city"      "year"      "month"     "sales"
[5] "volume"    "median"    "listings"  "inventory"
[9] "date"

> ##change months from numeric values
> data$month <- month.name[data$month]

> ##data summary
> str(data)
tibble [8,602 × 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    : chr [1:8602] "January" "February" "March" "April" ...
 $ 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 ...
```

After gaining a basic understanding of the dataset, I checked for missing values. The results are shown in Figure 2. After determining that the rows with missing values were useless to the analysis, I used na.omit to delete them before continuing. This is shown in Figure 3.

Figure 2: Checking for Missing Data

```
> ##check for missing values
> sapply(data, function(x)sum(is.na(x)))
     city      year     month     sales    volume    median  listings inventory      date
        0         0         0       568       568       616      1424      1467         0

> ##view sample of rows with missing data
> missing_data <- txhousing[is.na(txhousing$sales),]

> print(head(missing_data))
# A tibble: 6 × 9
  city             year month sales volume median listings inventory  date
  <chr>           <int> <int> <dbl>  <dbl>  <dbl>    <dbl>     <dbl> <dbl>
1 Brazoria County  2001    10    NA     NA     NA       NA        NA 2002.
2 Brazoria County  2003     1    NA     NA     NA       NA        NA 2003
3 Brazoria County  2003     2    NA     NA     NA       NA        NA 2003.
4 Brazoria County  2003     3    NA     NA     NA       NA        NA 2003.
5 Brazoria County  2003     4    NA     NA     NA       NA        NA 2003.
6 Brazoria County  2003     5    NA     NA     NA       NA        NA 2003.
```

Figure 3: Solution to Missing Data

```
> ##delete rows that have missing data
> CleanData <- na.omit(txhousing)

> ##check for missing values
> sapply(CleanData, function(x)sum(is.na(x)))
     city      year     month     sales    volume    median  listings inventory      date
        0         0         0         0         0         0         0         0         0
```

III. SUMMARY STATISTICS

Figure 4 represents a summary view of the data by the statistical measures of minimum, maximum, mean, median, 1st quartile median, and 3rd quartile median.

Figure 4: Summary Statistics

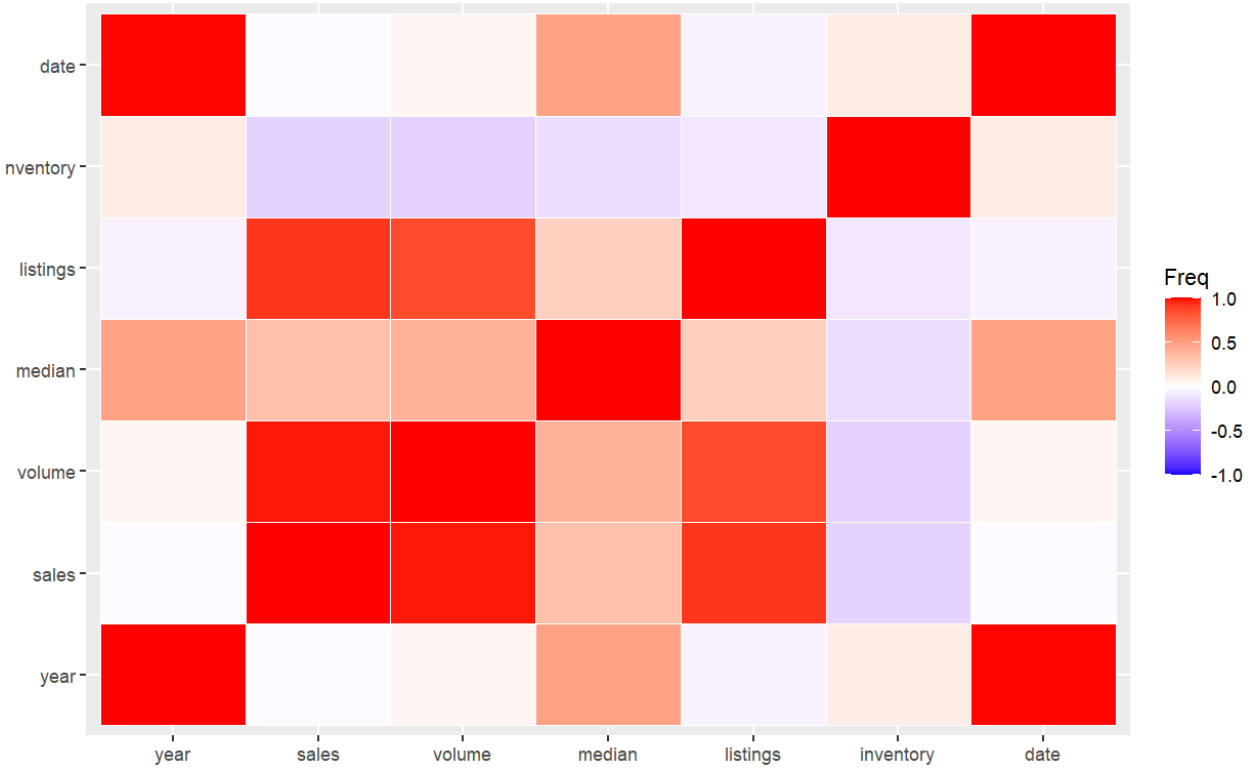
```
> summary(data)
     city                year          month
 Length:8602        Min.   :2000   Length:8602
 Class :character   1st Qu.:2003   Class :character
 Mode  :character   Median :2007   Mode  :character
                    Mean   :2007
                    3rd Qu.:2011
                    Max.   :2015

     sales             volume              median
 Min.   :   6.0   Min.   :8.350e+05   Min.   : 50000
 1st Qu.:  86.0   1st Qu.:1.084e+07   1st Qu.:100000
 Median : 169.0   Median :2.299e+07   Median :123800
 Mean   : 549.6   Mean   :1.069e+08   Mean   :128131
 3rd Qu.: 467.0   3rd Qu.:7.512e+07   3rd Qu.:150000
 Max.   :8945.0   Max.   :2.568e+09   Max.   :304200
 NA's   :568      NA's   :568         NA's   :616
    listings       inventory            date
 Min.   :    0   Min.   : 0.000   Min.   :2000
 1st Qu.:  682   1st Qu.: 4.900   1st Qu.:2004
 Median : 1283   Median : 6.200   Median :2008
 Mean   : 3217   Mean   : 7.175   Mean   :2008
 3rd Qu.: 2954   3rd Qu.: 8.150   3rd Qu.:2012
 Max.   :43107   Max.   :55.900   Max.   :2016
 NA's   :1424    NA's   :1467
```

Figure 5 is a correlation matrix to show the correlation between all numerical and integer values. Using the data from the correlation matrix, a correlation heatmap is shown in Figure 6 to better understand the correlations between each variable. Darker red cells represent a strong positive correlation, while darker blue cells represent a strong negative correlation.

Figure 5: Correlation Matrix

	year	sales	volume	median	listings	inventory	date
year	1.00000000	-0.01619670	0.04996410	0.4829000	-0.05405689	0.09757065	0.99800654
sales	-0.01619670	1.00000000	0.98080912	0.3350742	0.92139125	-0.19290223	-0.01513078
volume	0.04996410	0.98080912	1.00000000	0.4021101	0.86074997	-0.19426562	0.05115138
median	0.48289997	0.33507418	0.40211005	1.0000000	0.24560270	-0.14218928	0.48564392
listings	-0.05405689	0.92139125	0.86074997	0.2456027	1.00000000	-0.10019973	-0.05372499
inventory	0.09757065	-0.19290223	-0.19426562	-0.1421893	-0.10019973	1.00000000	0.09856982
date	0.99800654	-0.01513078	0.05115138	0.4856439	-0.05372499	0.09856982	1.00000000

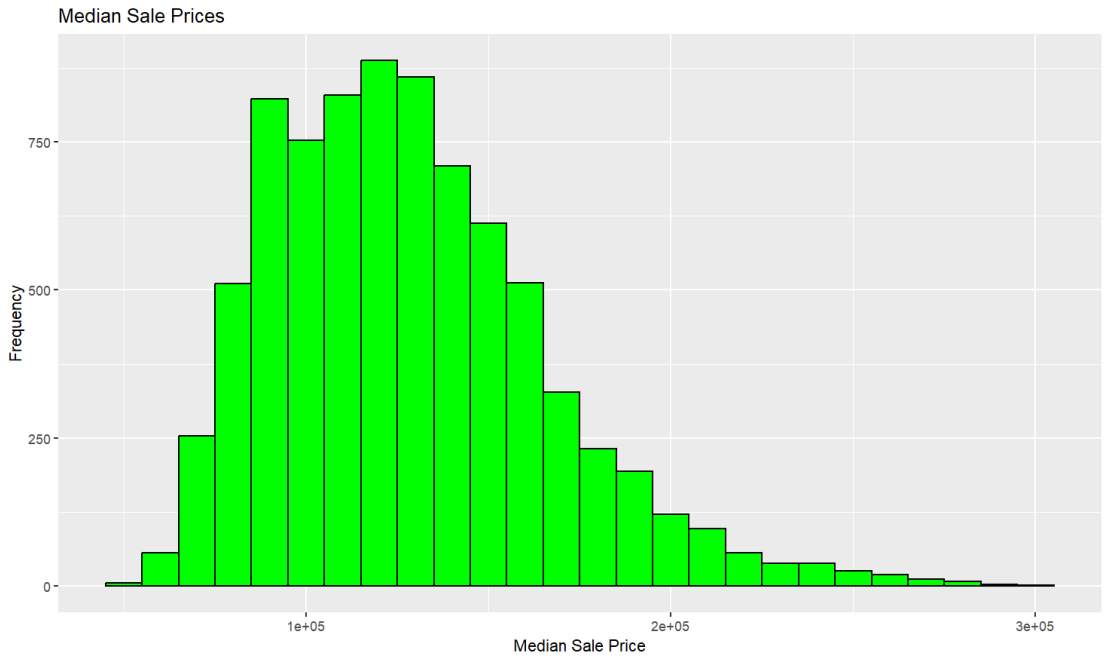
Figure 6: Correlation Heatmap



IV. DATASET GRAPHICAL EXPLORATION

The following univariate chart, Figure 7, explores the frequency of different median sales prices. This tells us how many sales occurred at different median price points. Based on the graph, we can see that the frequency of sales peaked when median prices were between $100k and $150k.

Figure 7: Median Sale Prices



In almost every industry, sales figures typically change based on the month with demand. In Figure 8, I sought to explore if the Texas housing market would operate in the same way since the Winter season does not get as cold there. This graphical analysis would suggest that there is not a huge difference in sales across different seasons, but between the months of August and December, sales decreased slightly.

Figure 8: Sales by Month

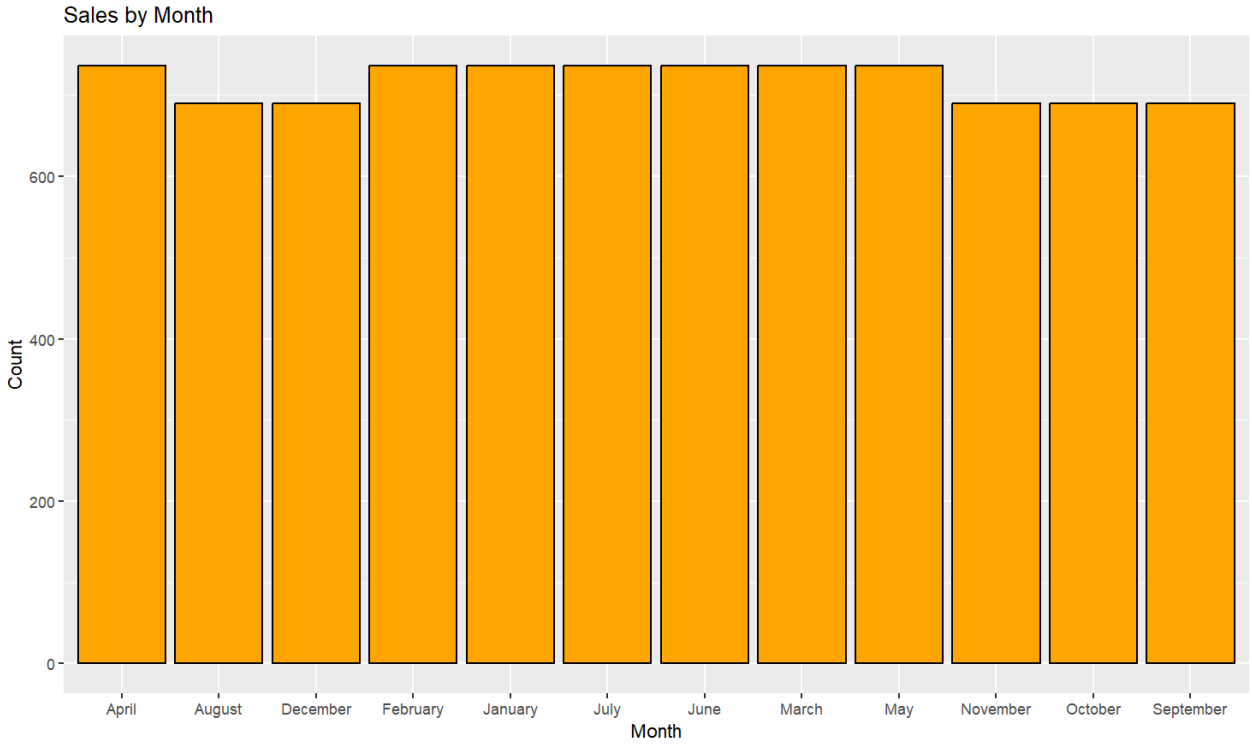


Figure 9 is an exploration of the number of sales in each city for each month. It is a univariate box plot that shows how the bulk of cities perform, then the outliers which would represent mostly just urban cities. This box plot shows that the majority of cities have a very low number of sales, but there are quite a few outliers above, most likely representing major cities across the huge state.

Figure 9: Box Plot of Sales

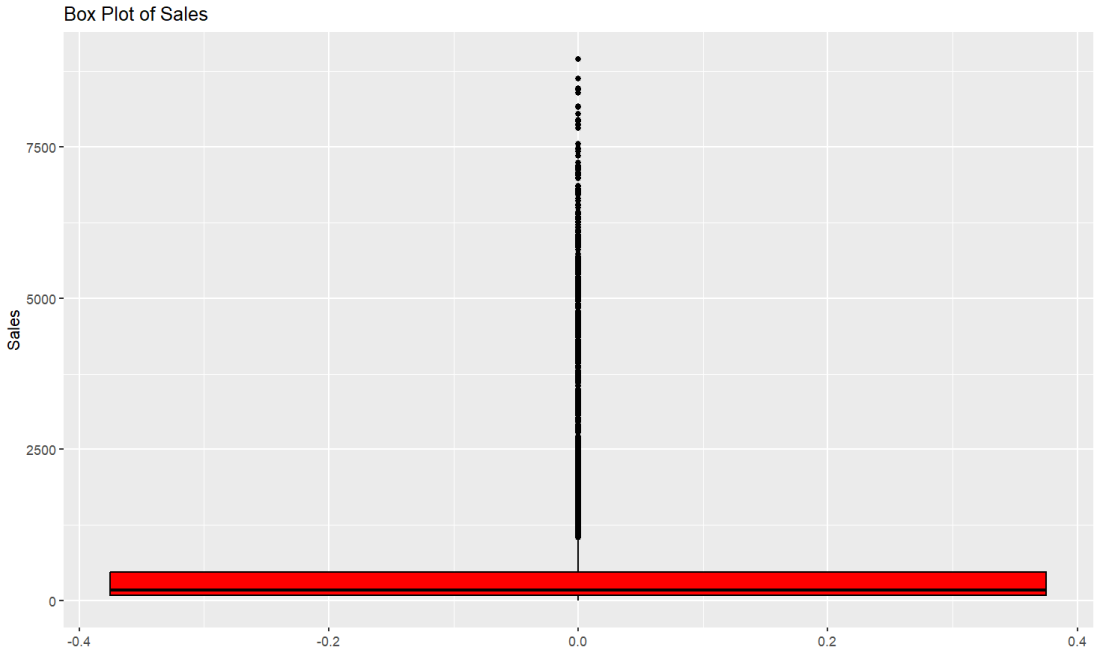
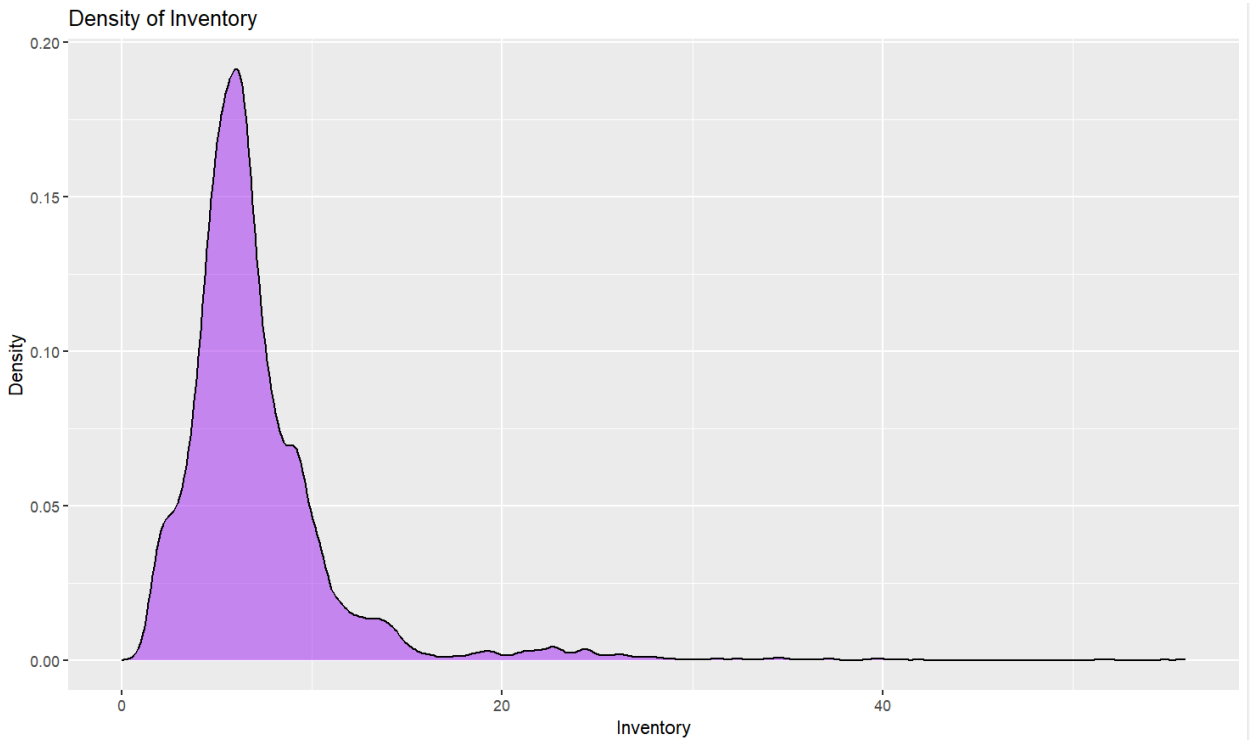


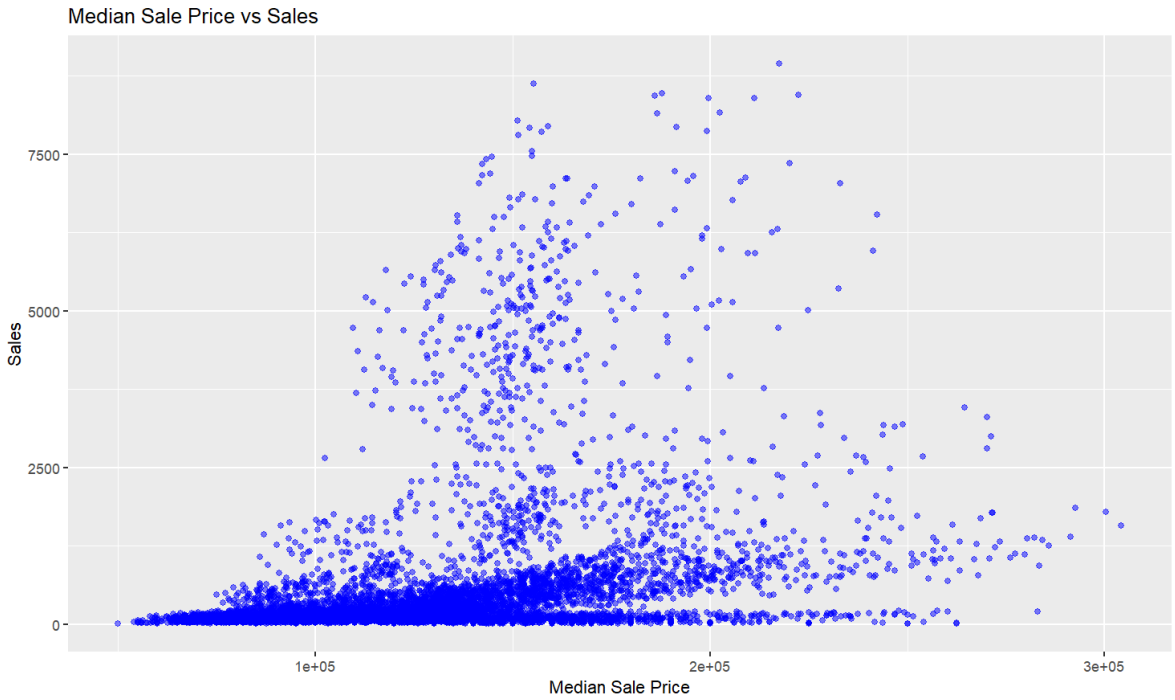
Figure 10 is a univariate density plot showing the density of the total amount of available inventory. Based on the graph, the majority of cities had less than 20 months' worth of inventory available at any given time.

Figure 10: Density of Inventory



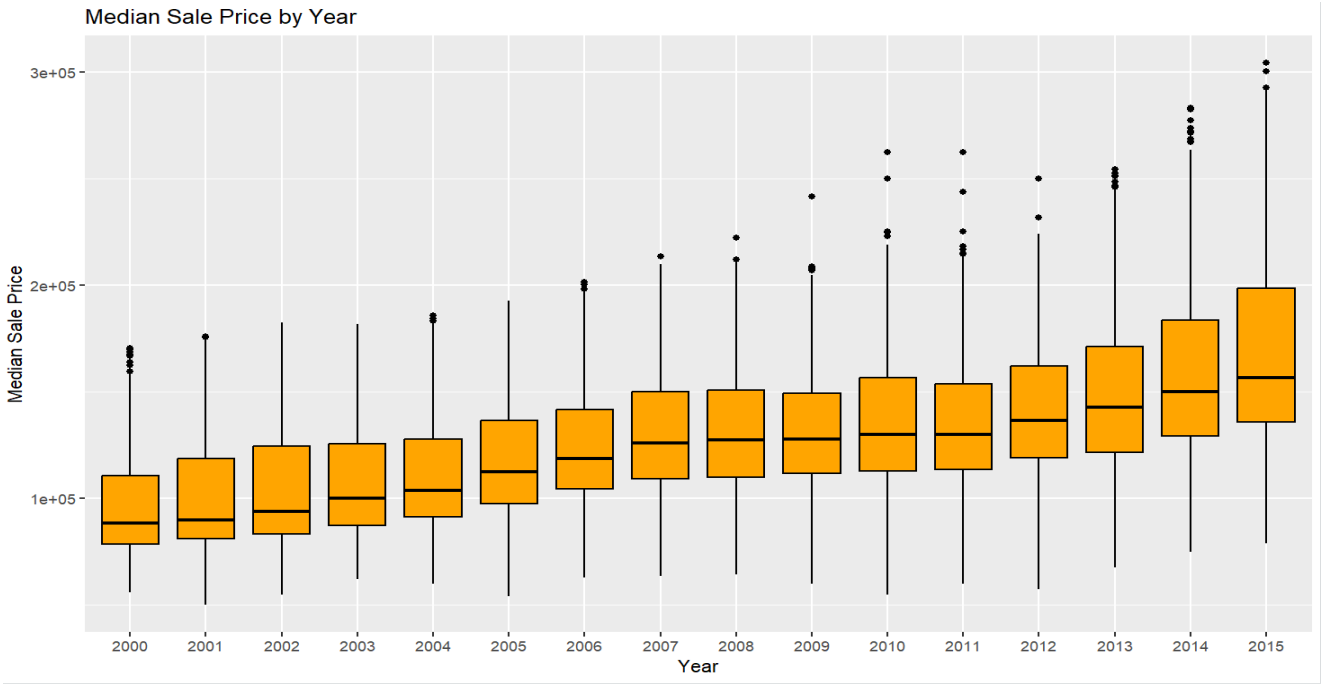
Bivariate plots allow a data scientist to explore the relationship between two variables. Figure 11 is a bivariate scatterplot exploring the relationship between the median sale price and total sales. This plot shows that the median sale price had little effect on the total number of sales in the majority of cities.

Figure 11: Median Sale Price vs Sales



I wanted to find out if there was an increase in the average market price for houses in Texas between 2000 and 2015. I used a bivariate box plot to explore this in Figure 12. The box plot reflects that there was a steady increase in median sale price across this time period.

Figure 12: Median Sale Price by Year



I used another bivariate box plot to understand how much the available inventory changed throughout this time period. This is shown in Figure 13. Based on the graph, the amount of available inventory did not change much throughout this time period, but there are some outliers. One example is in 2010, where there was probably rapid building of homes or just general vacancies.

Figure 13: Inventory by Year

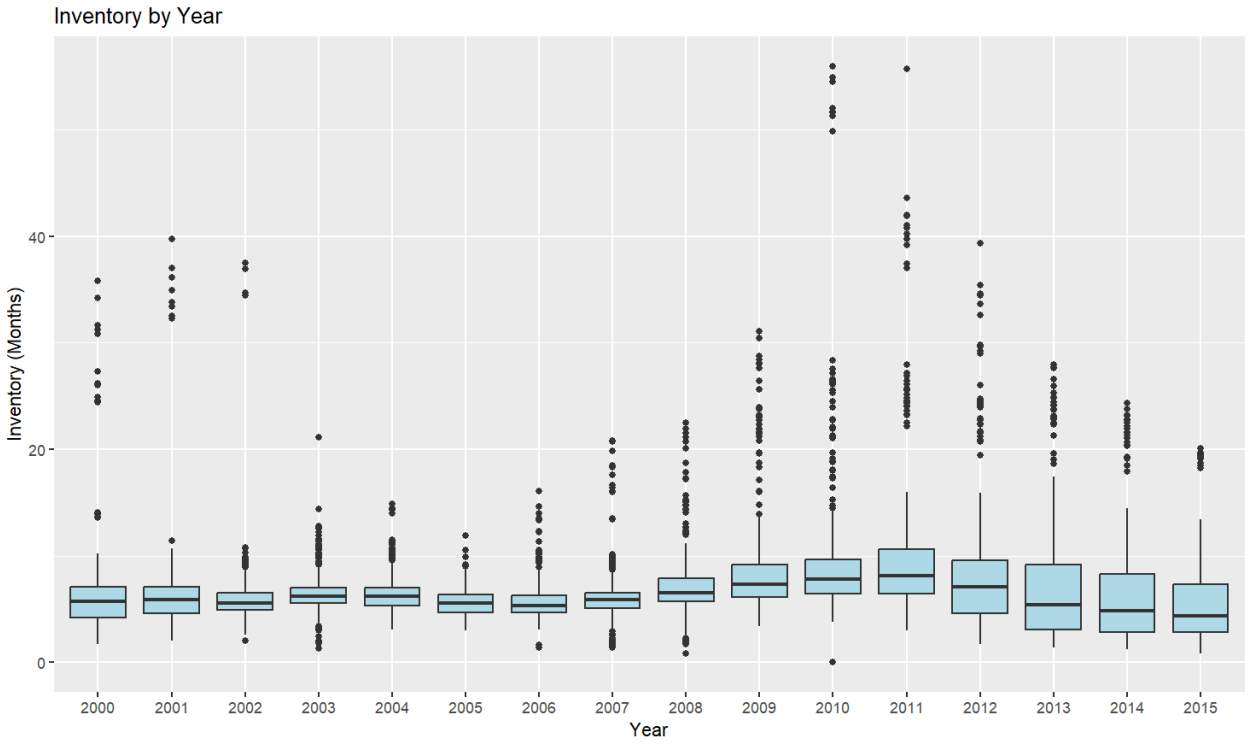


Figure 14 explores the relationship between the total number of sales and the monetary value of the total sales (volume). It is a bivariate scatterplot. There appears to be a very strong correlation between these two variables.

Figure 14: Volume vs Sales

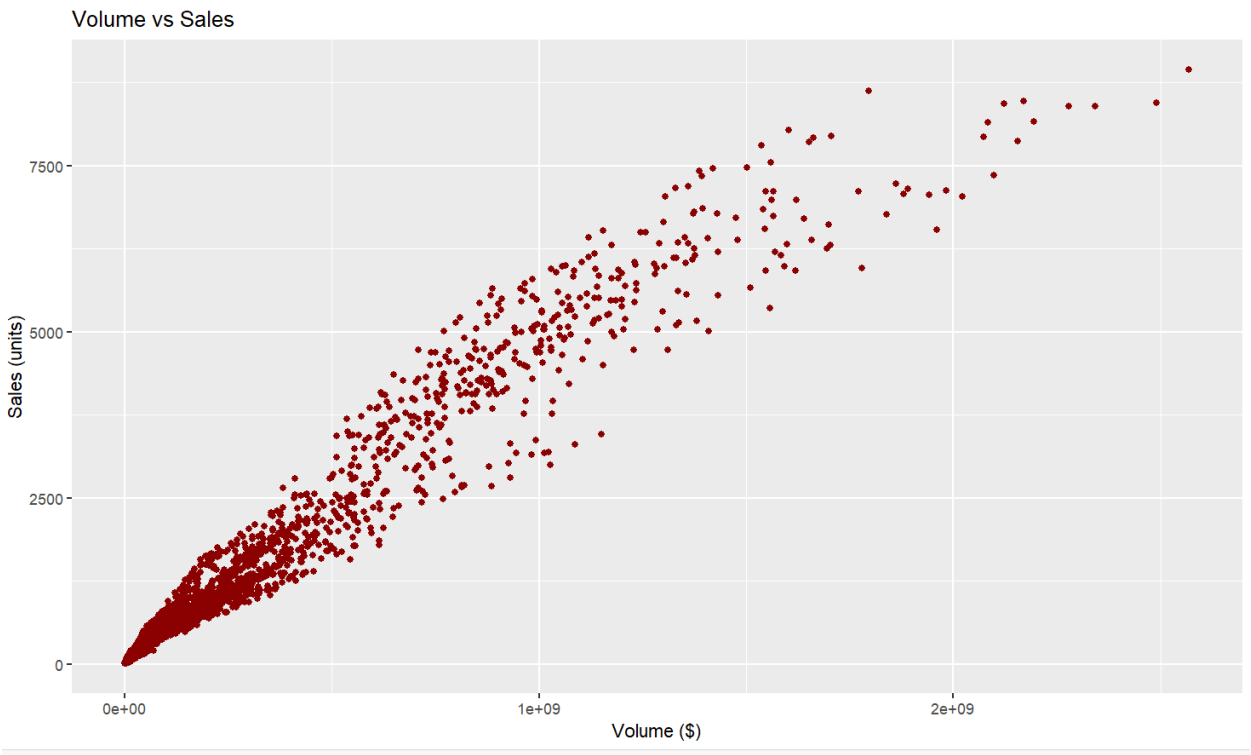
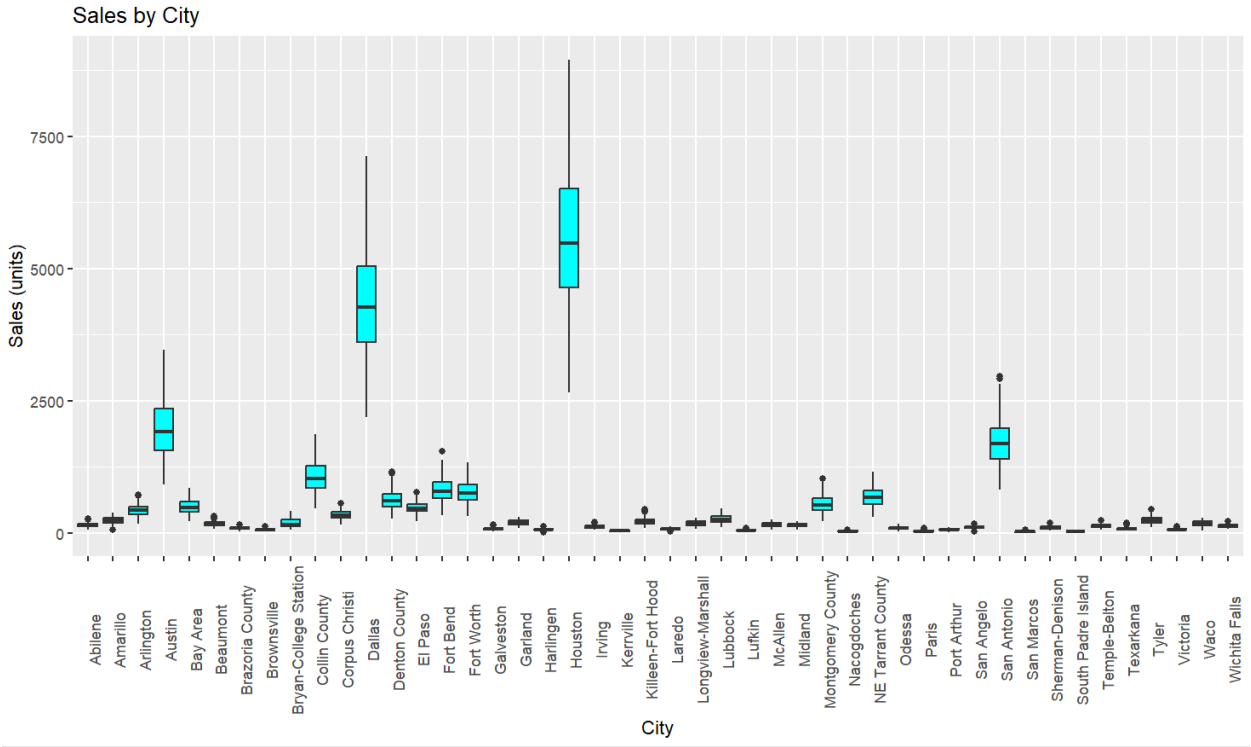


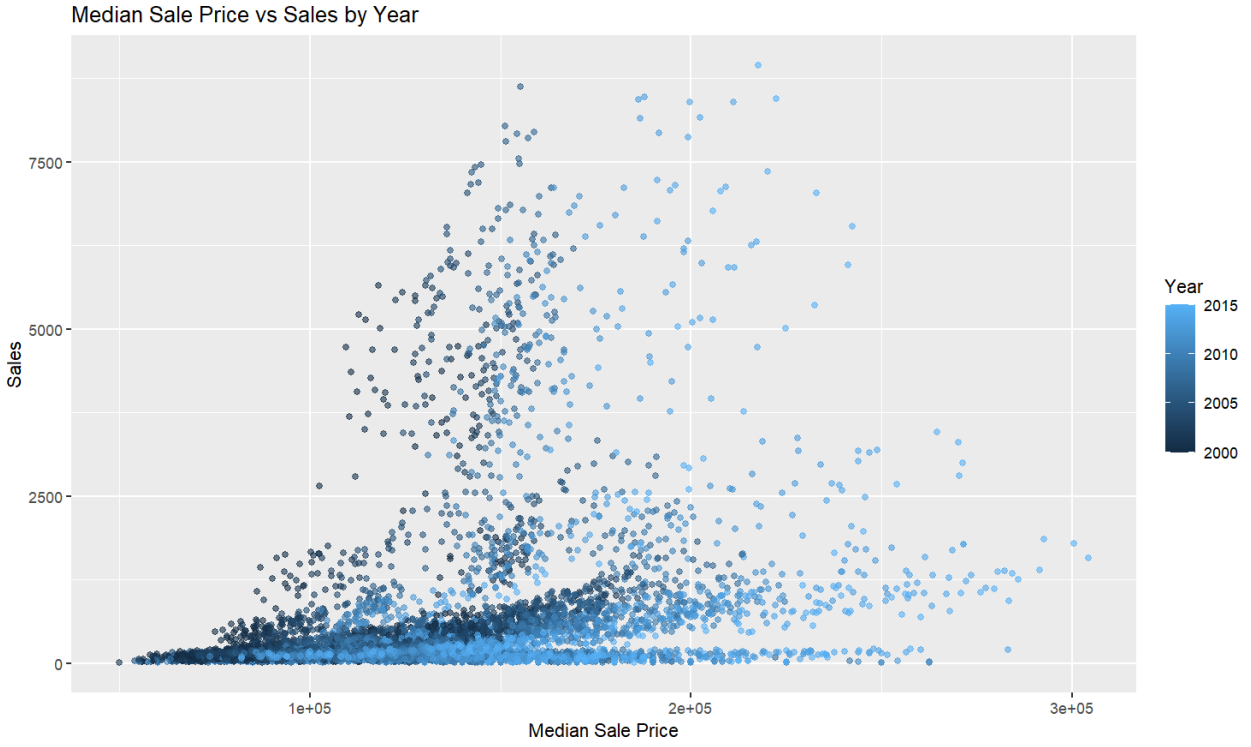
Figure 15 analyzes the total number of sales by city using a bivariate box plot. This box plot highlights how different the housing market is in Texas' big cities versus the majority of the state. It is clear that Dallas and Houston have a lot higher average number of sales.

Figure 15: Sales by City



Multivariate plots are useful at analyzing more than 2 variables within one plot. The correlation matrix and correlation heatmap were two multivariate plots used at the beginning. Figure 16 is a third multivariate plot. It shows the relationship between the median sale price and the number of sales, then also uses different shades of blue to represent different years. This graph reflects the same result as Figure 11, but offers more valuable insight. We can see from this graph that the median sale price increased as time continued, but it still does not appear that relationship between median sale price and total sales changed much.

Figure 16: Median Sale Price vs Sales by year



V. Findings

In conclusion, the Texas Housing dataset has a lot of valuable information about Texas' housing market between 2000 and 2015. Several key insights are clear. Firstly, the big cities account for a huge portion of the housing market sales during this time. Unsurprisingly, this means that the majority of small Texas cities had a small number of house sales. This analysis also suggests that there is a very high demand for homes between the $100k and $150k price range (figure 7). Unfortunately for Texas home buyers, Figure 12 shows that home prices are steadily increasing so they will likely have to open up to spending more on houses. Figure 8 shows that there is not a huge difference in total sales throughout different months, but August-December had a little less business. Figure 15 shows that there is definitely a correlation between volume and sales. Based on the heatmap, a few different variables have a strong correlation with total sales, the target variable of this whole dataset. Volume and total listings (homes on the market) are clearly the biggest influencers on the number of sales. Considering this relationship, it is unsurprising that there is also a strong correlation between the total volume and number of listings.