EXPLORATIVE ANALYSIS OF TEXAS REAL ESTATE MARKET

2. INDICATE THE "realestate_texas.csv" VARIABLE'S TYPES:

- **city**: qualitative nominal;
- **year**: qualitative continuous, (treated like ordinal);
- **month**: qualitative nominal (cyclic), codified in numbers;
- **sales**: quantitative discrete on ratio scale;
- **volume**: quantitative continuous on ratio scale;
- **median_price**: quantitative continuous on ratio scale;
- **listings**: quantitative discrete on ratio scale;
- **months_inventory**: quantitative continuous on ratio scale;

3. POSITION, VARIABILITY AND SHAPE INDEXES OF ALL THE VARIABLES (when possible):

- City:

	ni1	fi1	Ni1	Fi1
Beaumont	60	0.25	60	0.25
Bryan-College Station	60	0.25	120	0.50
Tyler	60	0.25	180	0.75
Wichita Falls	60	0.25	240	1.00

- Year:

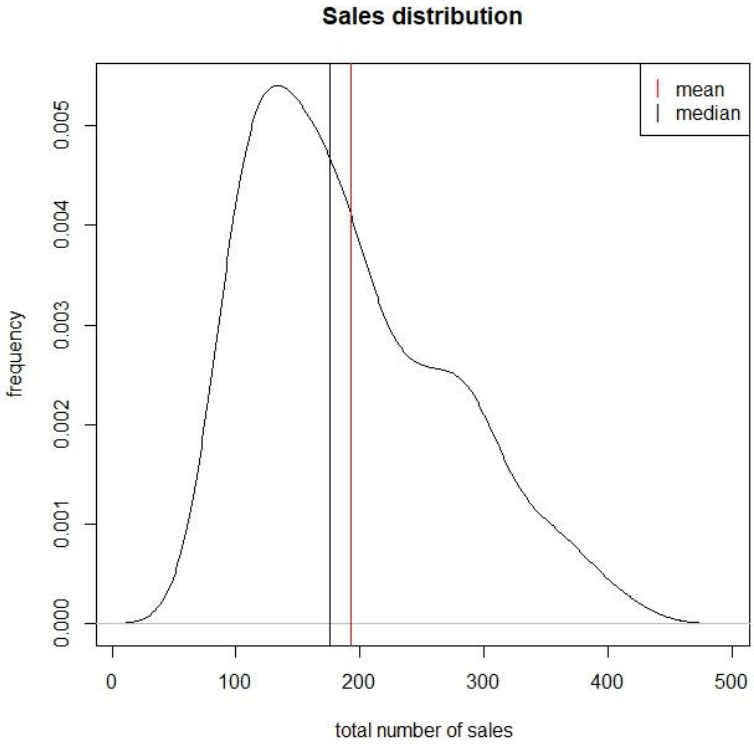
	ni2	fi2	Ni2	Fi2
2010	48	0.2	48	0.2
2011	48	0.2	96	0.4
2012	48	0.2	144	0.6
2013	48	0.2	192	0.8
2014	48	0.2	240	1.0

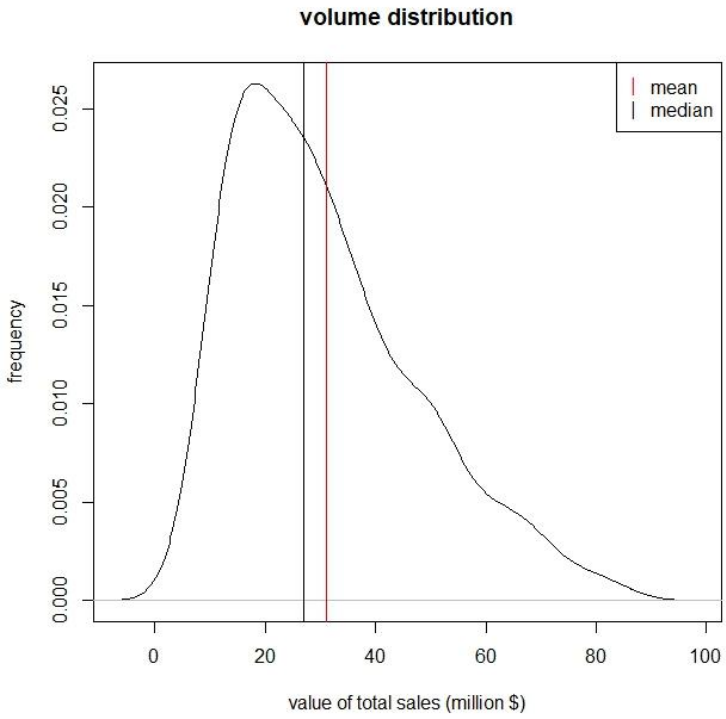
- Month:

	ni3	fi3	Ni3	Fi3
1	20	0.08333333	20	0.08333333
2	20	0.08333333	40	0.16666667
3	20	0.08333333	60	0.25000000
4	20	0.08333333	80	0.33333333
5	20	0.08333333	100	0.41666667
6	20	0.08333333	120	0.50000000
7	20	0.08333333	140	0.58333333
8	20	0.08333333	160	0.66666667
9	20	0.08333333	180	0.75000000
10	20	0.08333333	200	0.83333333
11	20	0.08333333	220	0.91666667
12	20	0.08333333	240	1.00000000

The data shown regard four different Texas cities (equally represented in terms of numerosity) within a time of five years (2010-2014), each month of the year. The data frame has 240 rows, thus each city occupies 60 of them (240/4). Each year has 48 rows (240/5) and each month has 20 (240/12) equally spread amongst 5 years (20/5 = 4 per year). Being these variables qualitatives the analysis of the position, variability, and shape indexes was not performed.

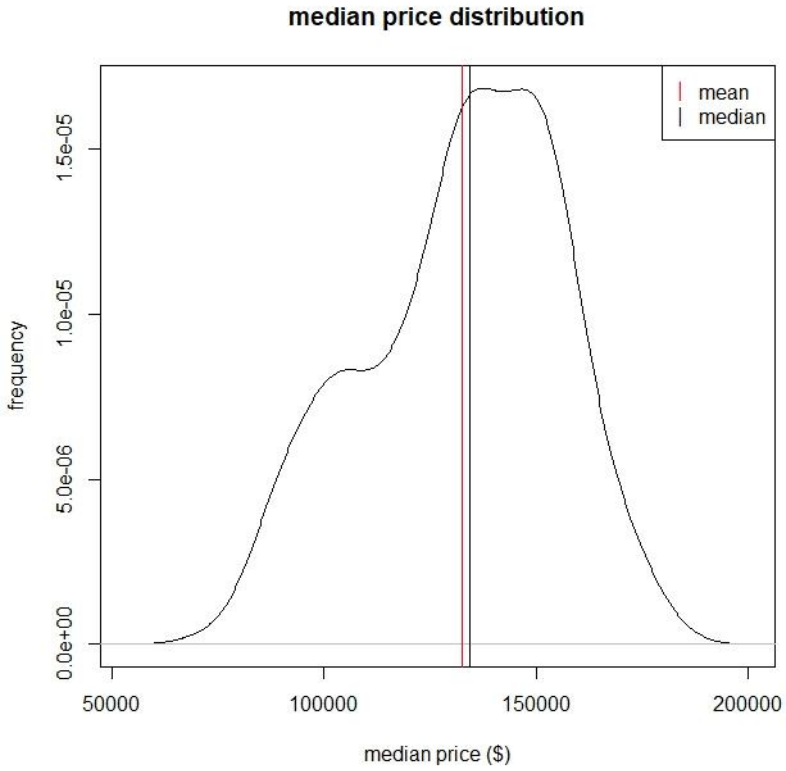
- "sales": the mean is 192.29 and the median is 175.5. The range is 79-423 and the IQR is 120 (total number of sales). Moreover, the coefficient of variation is 41.42. μ_3 is 0.72 and μ_4 is -1.22, therefore the variable has a positive-asymmetric, platykurtic distribution.

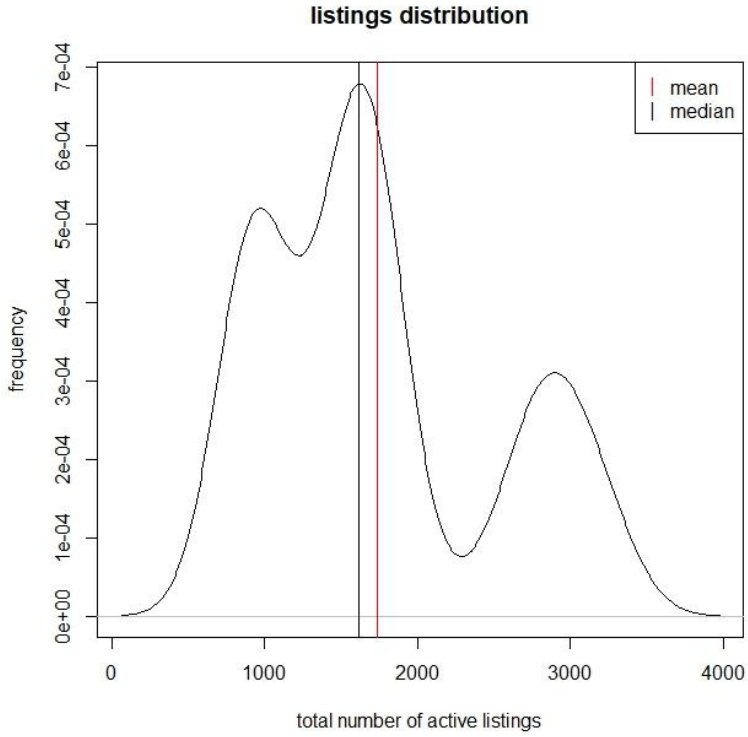




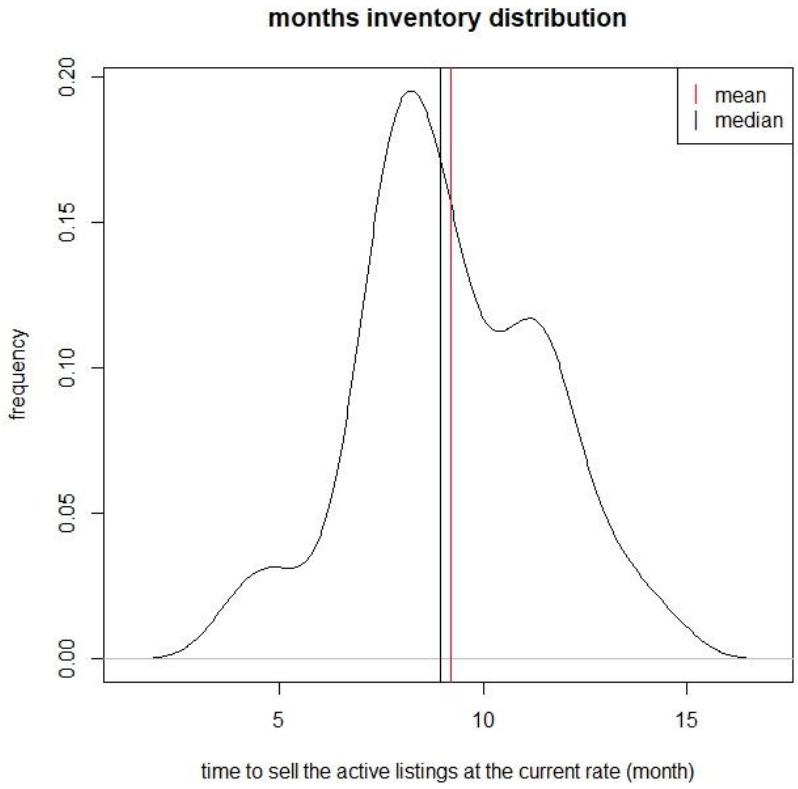
- "volume": the mean is 31 and the median is 27.06. The range is 8.17-83.55 and the IQR is 23.23 (million $). In addition, its coefficient of variation is 53.70. μ_3 is 0.88 and μ_4 is 0.17, thus the variable has a positive-asymmetric, leptokurtic distribution.

- “median_price”: the mean is 132665.42 and the median is 134500. The range is 73800-180000 and the IQR is 32750 ($). Additionally, the coefficient of variation is 17.08. μ_3 is -0.36 and μ_4 is -0.62, hence the distribution is negative-asymmetric and platykurtic.





- “listings”: the mean is 1738.02 and the median is 1618.5. The range is 743-3296 and the IQR is 1029.5 (number of active listings). Moreover, the coefficient of variation is 43.31. μ_3 is 0.65 and μ_4 is -0.79, therefore the variable has a positive-asymmetric, platykurtic distribution, even though it is hardly comparable to a Gaussian curve.



• "months_inventory": the mean is 9.19 and the median is 8.95. The range is 3.4-14.9 and the IQR is 3.15 (months). Furthermore, the coefficient of variation is 25.06. μ_3 is 0.04 and μ_4 is -0.17, thus the variable distribution is positive-asymmetric and platykurtic.

4. WHICH IS THE MOST VARYING VARIABLE AND WHICH IS THE MOST ASYMMETRIC:

Looking at the Coefficients of variation related to the five quantitative variables (the remaining three are equally distributed) it is easy to notice how "volume" (the 5th) is more spread around its mean than the others.

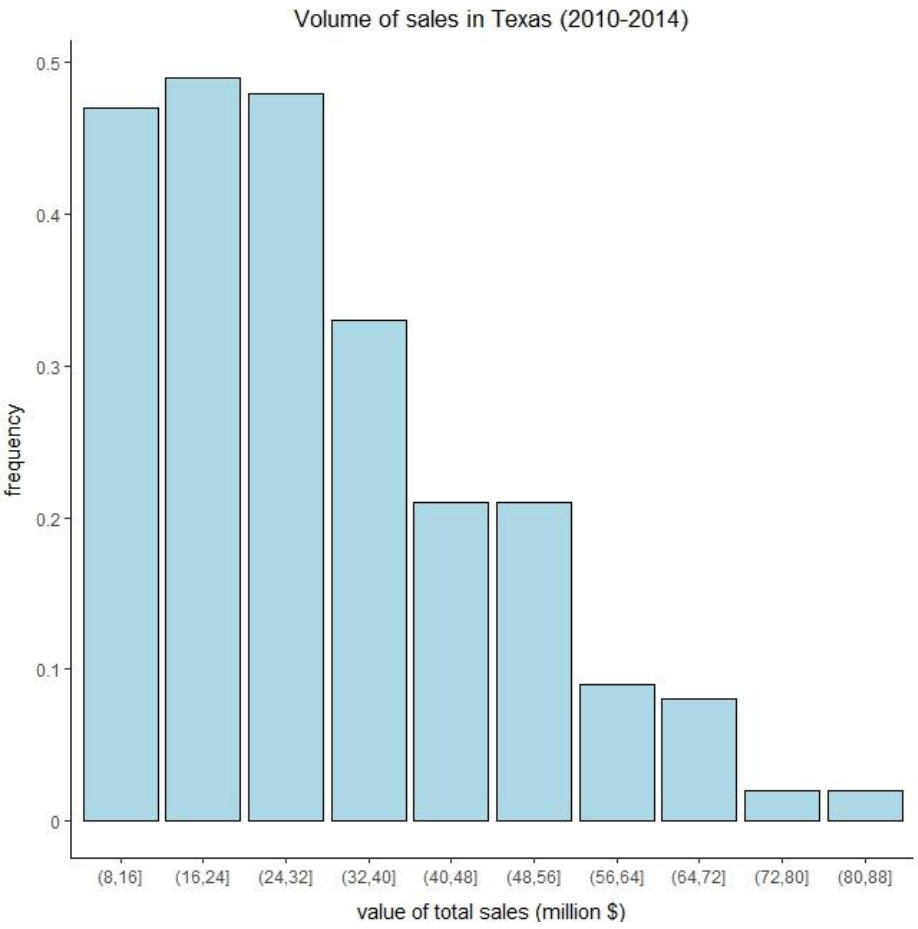
```
> CVs
         cv4      cv5      cv6      cv7      cv8
[1,] 41.42203 53.70536 17.08218 43.30833 25.06031
```

Considering the Fisher's Skewness Indexes emerges that "volume" is again the most asymmetric variable in absolute terms (without evaluating the sign).

```
> MU3s
        mu3_4    mu3_5      mu3_6     mu3_7      mu3_8
[1,] 0.718104 0.884742 -0.3645529 0.6494982 0.04097527
```

5. FREQUENCY TABLE, BAR GRAPH, AND GINI'S INDEX OF "volume" VARIABLE:

Subdividing the variable "volume" into equally long classes emerges that the vast part of sales in the Texas market consisted of values between 8 and 40 million dollars (73.7%, table and graph below).



	ni	fi	Ni	Fi
(8,16]	47	0.195833333	47	0.1958333
(16,24]	49	0.204166667	96	0.4000000
(24,32]	48	0.200000000	144	0.6000000
(32,40]	33	0.137500000	177	0.7375000
(40,48]	21	0.087500000	198	0.8250000
(48,56]	21	0.087500000	219	0.9125000
(56,64]	9	0.037500000	228	0.9500000
(64,72]	8	0.033333333	236	0.9833333
(72,80]	2	0.008333333	238	0.9916667
(80,88]	2	0.008333333	240	1.0000000

For the same reason, larger sales in terms of money (40-88 million dollars), were less influential on the total volume of money, being less frequent (26.3%, table and graph above). The Gini Index was calculated as 0.94, certifying the variable as highly heterogeneous.

6. GUESS THE GINI INDEX FOR THE VARIABLE “city”:

Considering that the city variable presents this distribution of absolute and relative frequencies:

	ni	fi	Ni	Fi
Beaumont	60	0.25	60	0.25
Bryan-College Station	60	0.25	120	0.50
Tyler	60	0.25	180	0.75
Wichita Falls	60	0.25	240	1.00

It’s easy to determine that its Gini Index is equal to one, which indeed represents maximal heterogeneity or equidistribution.

7. PROBABILITY CALCULATIONS:

Utilizing the classic meaning of probability (favorable cases/total cases) turns out that there is a 0.25 probability of getting the value “Beaumont” of the city variable.

```
> nrow(re_texas[city == "Beaumont",])/nrow(re_texas)
[1] 0.25
```

In the same way, the probability of obtaining a specific value of the variable “month”, July in this case, is 0.08333333.

```
> nrow(re_texas[month == 7,])/nrow(re_texas)
[1] 0.08333333
```

Lastly, considering the month “December” and the year “2012” together, they have a 0.01666667 probability of occurring.

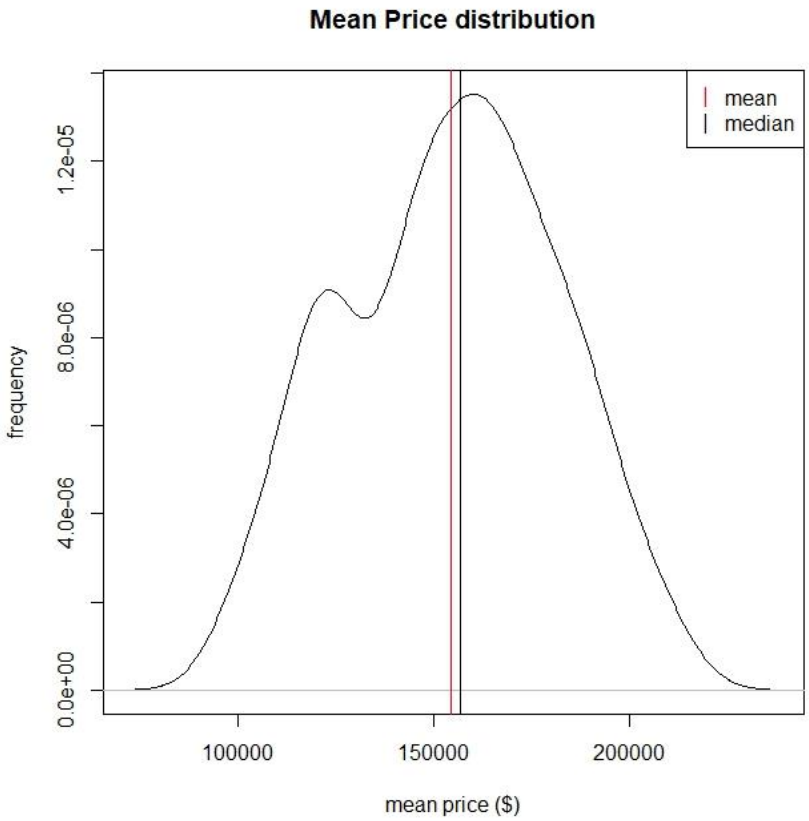
```
> nrow(re_texas[c(month == 12 & year == 2012),])/nrow(re_texas)
[1] 0.01666667
```

8. MEAN PRICE

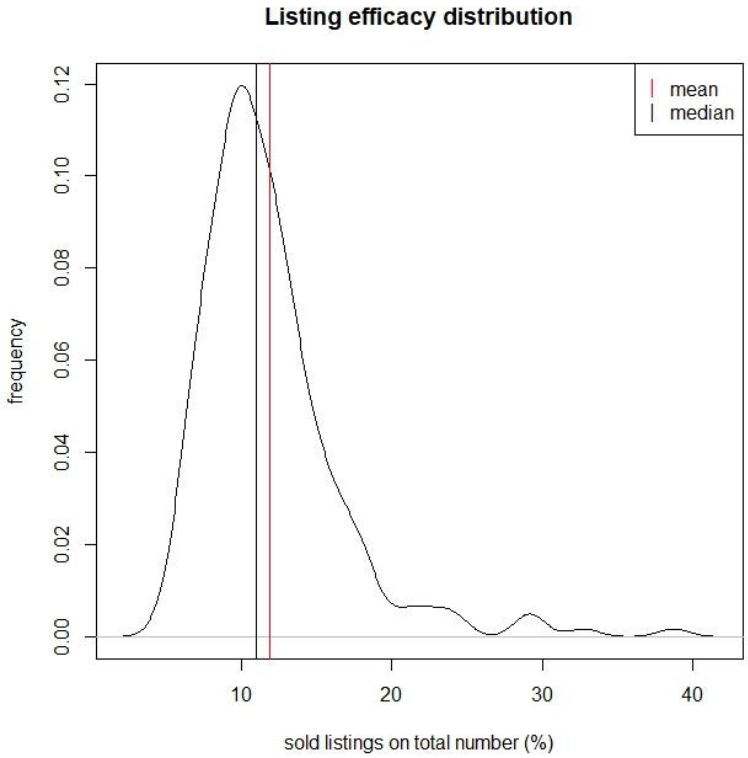
The mean price was calculated as the total value of sales on the total number of sales multiplied by 1000000 (to express it in dollars).

```
re_texas$mean_price = volume/sales*1000000
```

The mean is 154320.36 and the median is 156588.48. The range is 97010-213234 and the IQR is 40976.21 ($). In addition, the coefficient of variation is 17.59. μ_3 is -0.069 and μ_4 is -0.78, therefore the variable has a negative-asymmetric, platykurtic distribution.



9. DETERMINE THE EFFICACY OF LISTINGS



The efficacy of listings was calculated as:

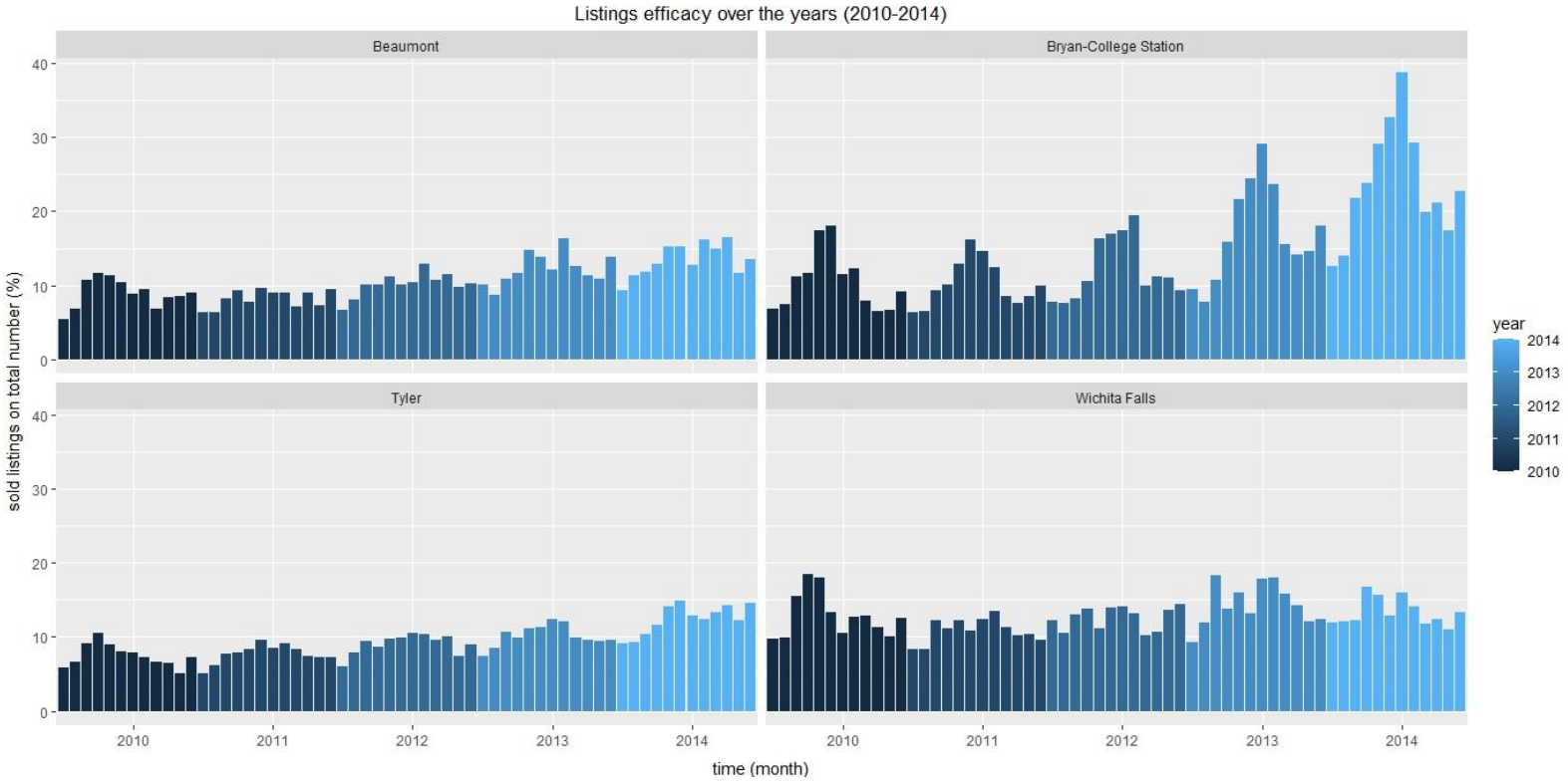
```
re_texas$list_efficacy = sales/listings*100
```

Thus considering the total number of sales in a month divided by the number of active announcements %.

The mean is 11.87 and the median is 10.96. The range is 5.01-38.71 and the IQR is 4.51 (number of listings sold in a month on the total (%)). The obtained coefficient of variation is 39.49. μ_3 is 2.09 and μ_4 is 6.88, hence the distribution is positive-asymmetric and leptokurtic.

To take into account the listings efficacy over the years *per city*, the latter variable was added through the *facet_wrap(~city)* command. The below panel shows how in each city the percentage of listings sold starts to increase at the beginning of the year and decreases again during the second half of each year (indicatively). Therefore, it looks like late Spring - Summer is the most efficacious period for selling real estate in Texas.

Examining each city per se, it seems that in two out of four towns, there is an increment over the 5 years, more or less pronounced (Bryan-College Station presents a more evident increase than Tyler). On the other side, Beaumont's efficacy initially diminishes to subsequently grow from 2011 to 2014 and Wichita Falls presents a more unorganized pattern. Generally speaking, Bryan-College Station has the highest efficacy followed by Wichita Falls, which precedes Beaumont, and Tyler.



10. DPLYR: MEAN AND STANDARD DEVIATION OF "median_price" & "mean_price":

Of "median_price" per every year, city, and month:

```
# A tibble: 5 x 3
   year   media deviazione_standard
  <int>   <dbl>               <dbl>
1  2010 130192.              21822.
2  2011 127854.              21318.
3  2012 130077.              21432.
4  2013 135723.              21708.
5  2014 139481.              25625.
```

```
# A tibble: 4 x 3
  city                   media deviazione_standard
  <chr>                  <dbl>               <dbl>
1 Beaumont             129988.              10105.
2 Bryan-College Station 157488.              8852.
3 Tyler                141442.              9337.
4 Wichita Falls        101743.              11320.
```

```
# A tibble: 12 x 3
   month  media deviazione_standard
   <int>  <dbl>               <dbl>
 1     1 124250              25151.
 2     2 130075              22823.
 3     3 127415              23442.
 4     4 131490              21458.
 5     5 134485              18796.
 6     6 137620              19231.
 7     7 134750              21945.
 8     8 136675              22488.
 9     9 134040              24344.
10    10 133480              26358.
11    11 134305              24691.
12    12 133400              22810.
```

Of "mean_price" per every year, city, and month:

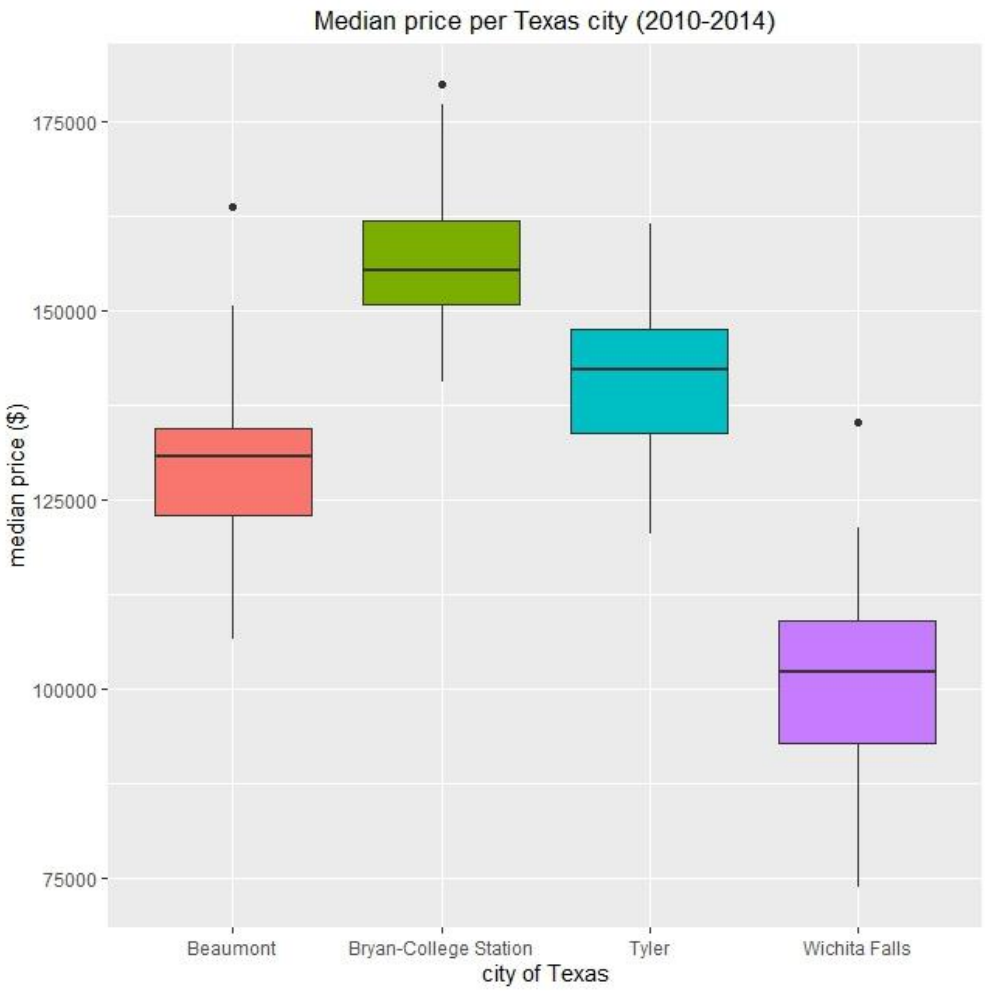
```
# A tibble: 5 x 3
   year   media deviazione_standard
  <int>   <dbl>               <dbl>
1  2010 150189.              23280.
2  2011 148251.              24938.
3  2012 150899.              26438.
4  2013 158705.              26524.
5  2014 163559.              31741.
```

```
# A tibble: 4 x 3
  city                   media deviazione_standard
  <chr>                  <dbl>               <dbl>
1 Beaumont             146640.              11232.
2 Bryan-College Station 183534.              15149.
3 Tyler                167677.              12351.
4 Wichita Falls        119430.              11398.
```

```
# A tibble: 12 x 3
   month   media deviazione_standard
   <int>   <dbl>               <dbl>
 1     1 145640.              29819.
 2     2 148840.              25120.
 3     3 151137.              23238.
 4     4 151461.              26174.
 5     5 158235.              25787.
 6     6 161546.              23470.
 7     7 156881.              27220.
 8     8 156456.              28253.
 9     9 156522.              29669.
10    10 155897.              32527.
11    11 154233.              29685.
12    12 154996.              27009.
```

11. GRAPHS:

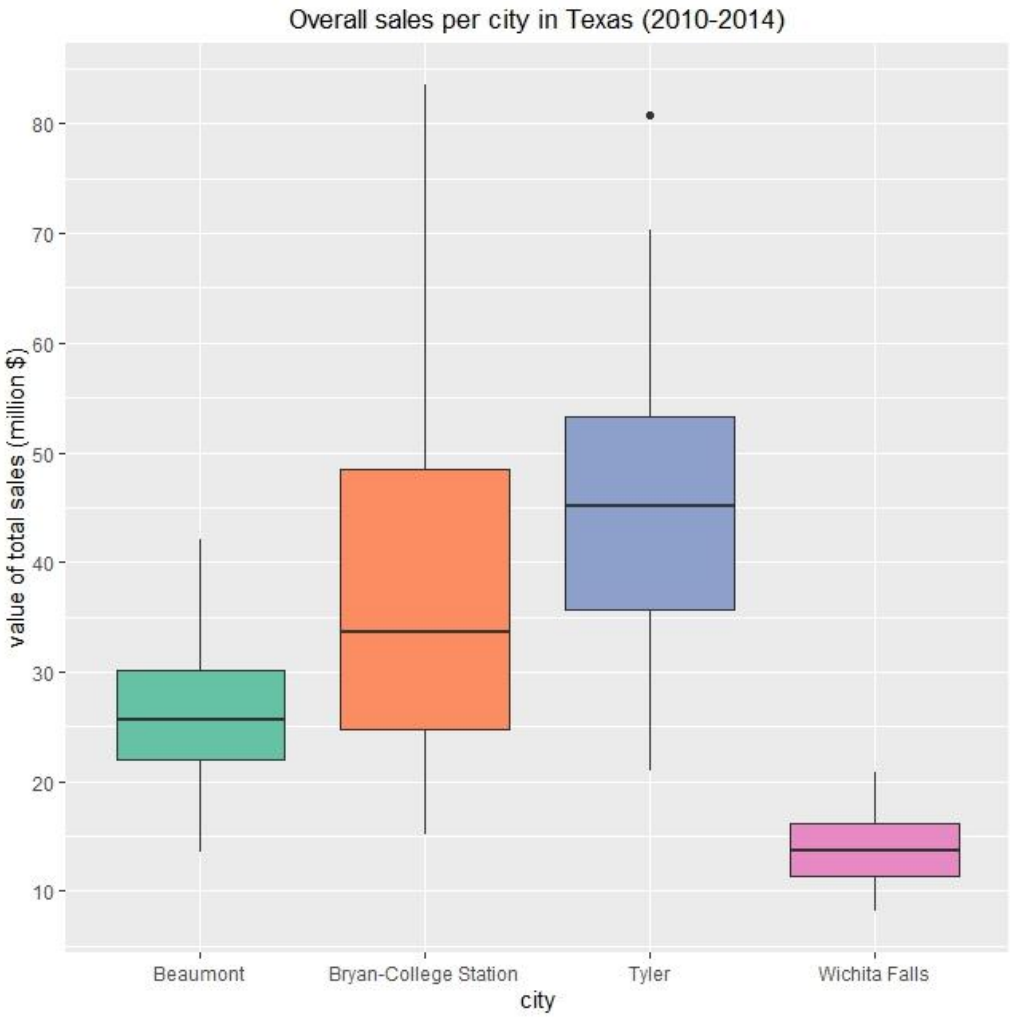
Median price per Texas city :



The graph above shows how the city with the higher median price is Bryan-College Station followed by Tyler, Beaumont, and Wichita Falls in this order. To add some information about the variability of "median_price" the *dplyr* package can be useful: looking at the table below, it is unambiguous how the median price variability is wider in Wichita Falls and progressively decreases in Beaumont, Tyler, and Bryan-College Station.

```
# A tibble: 4 x 3
  city                mediana deviazione_standard
  <chr>                 <dbl>               <dbl>
1 Beaumont             130750              10105.
2 Bryan-College Station 155400              8852.
3 Tyler                142200              9337.
4 Wichita Falls        102300              11320.
```

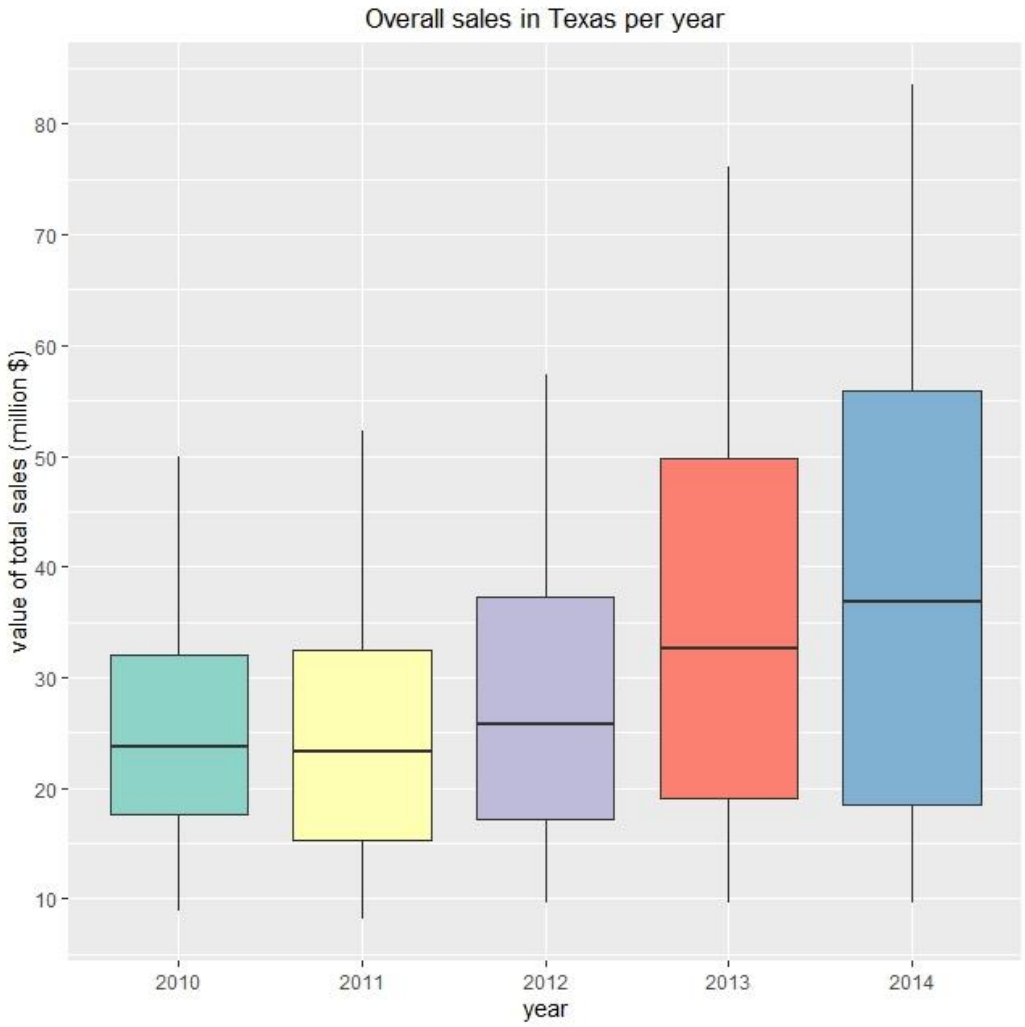
Overall sales per city in Texas (2010-2014)



The upper graph displays how the variability over the years of "volume" is lower in Wichita Falls and Beaumont compared to the other two cities, especially for what concerns the former. Moreover, despite Bryan-College Station has the variable highest value due to its more pronounced variability, the median of Tyler is higher (numerically represented by the table below).

```
# A tibble: 4 x 3
  city                mediana deviazione_standard
  <chr>                 <dbl>               <dbl>
1 Beaumont               25.6                6.97
2 Bryan-College Station  33.6               17.2
3 Tyler                  45.1               13.1
4 Wichita Falls          13.7                3.24
```

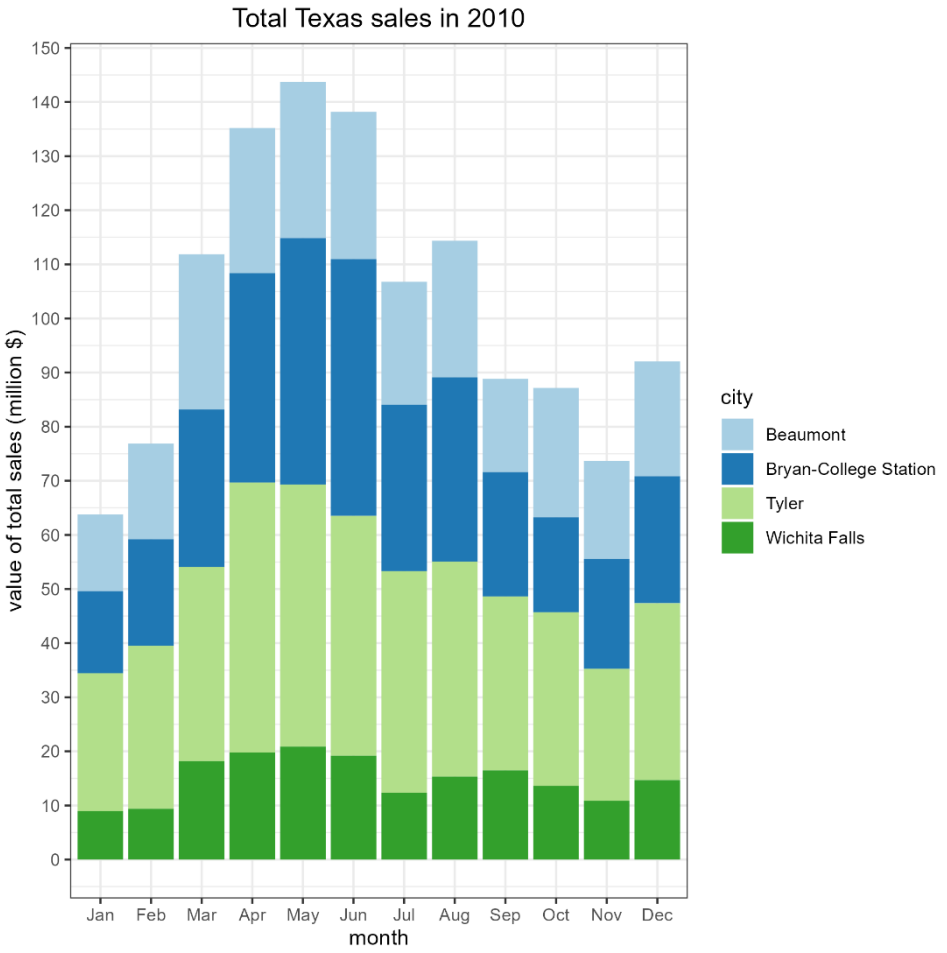
Overall sales in Texas per year



This representation better describes that from 2010 to 2014, there is a progressive increase in the median of the value of total sales and its variability (except for the 2010 to 2011 transition in which there is a slight decrease in the median value, quantified by the table below).

```
# A tibble: 5 x 3
  year  mediana deviazione_standard
  <fct>   <dbl>                <dbl>
1 2010     23.7                10.8
2 2011     23.2                12.2
3 2012     25.8                14.5
4 2013     32.7                17.9
5 2014     36.8                21.2
```

Total Texas sales per month, considering each year separately

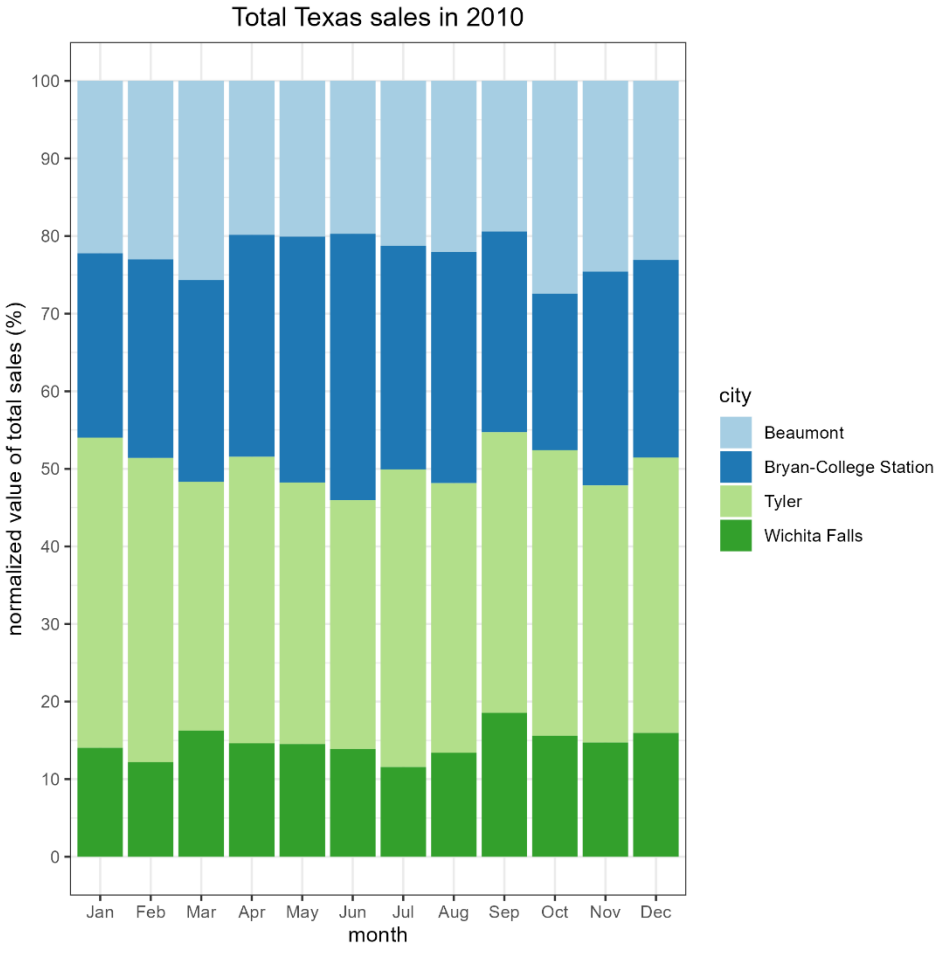


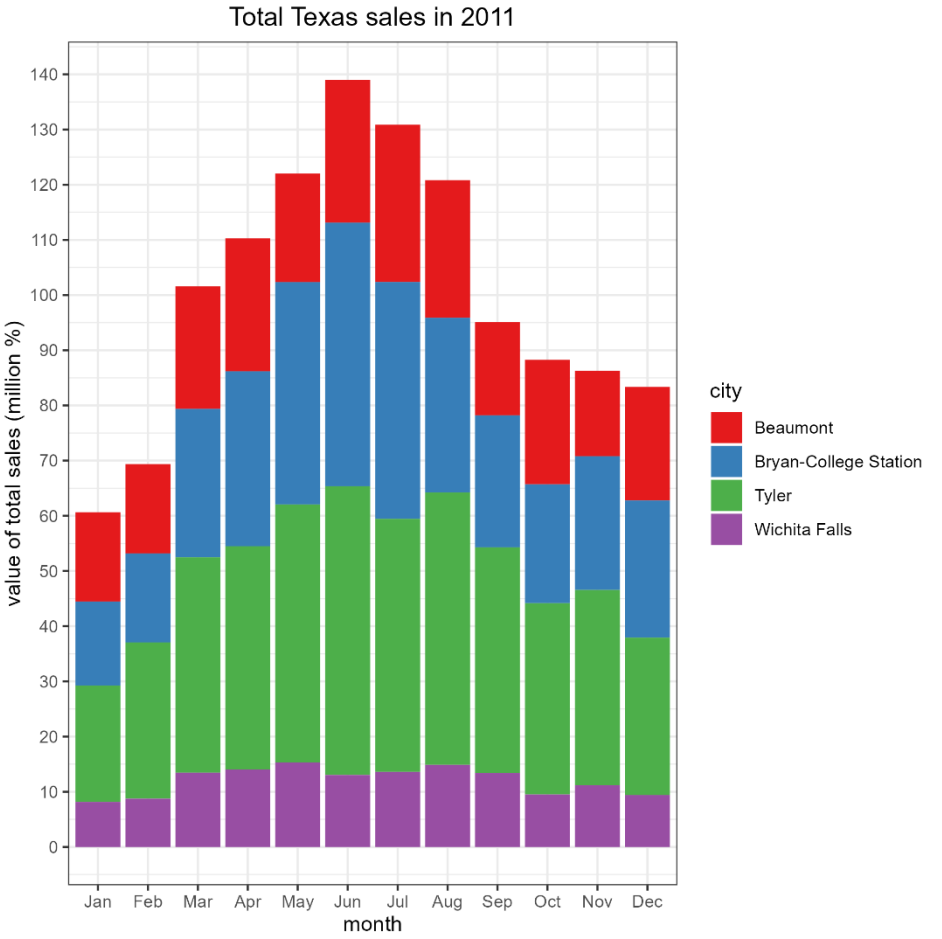
Taking into account the "volume" variable exclusively in 2010, it can be seen how Tyler city is characterized by the highest values, followed by Bryan-College Station, Beaumont, and Wichita Falls in every month of the year. In addition to that, the graph better shows how March, April, May, June, and August are the months in which the overall value of sales (four cities together) reaches its maximum, proving how Spring-Summer represents the best period of the year for the real estate market as was stated analyzing "list_eficacy" variable (graph on the left).

On the other hand, the normalized comparison displays how the proportions of the variable changed only a little during the different months of 2010, being represented by the following percentages:

1) Beaumont ~ 20-25 %;
2) Bryan-College Station ~ 20-30 %;
3) Tyler ~ 35-40 %;
4) Wichita Falls ~ 10-20 %;

These data confirm that Tyler owns the most consistent part of sales value, followed by Bryan-College Staton, Beaumont, and Wichita Falls (graph on the right).



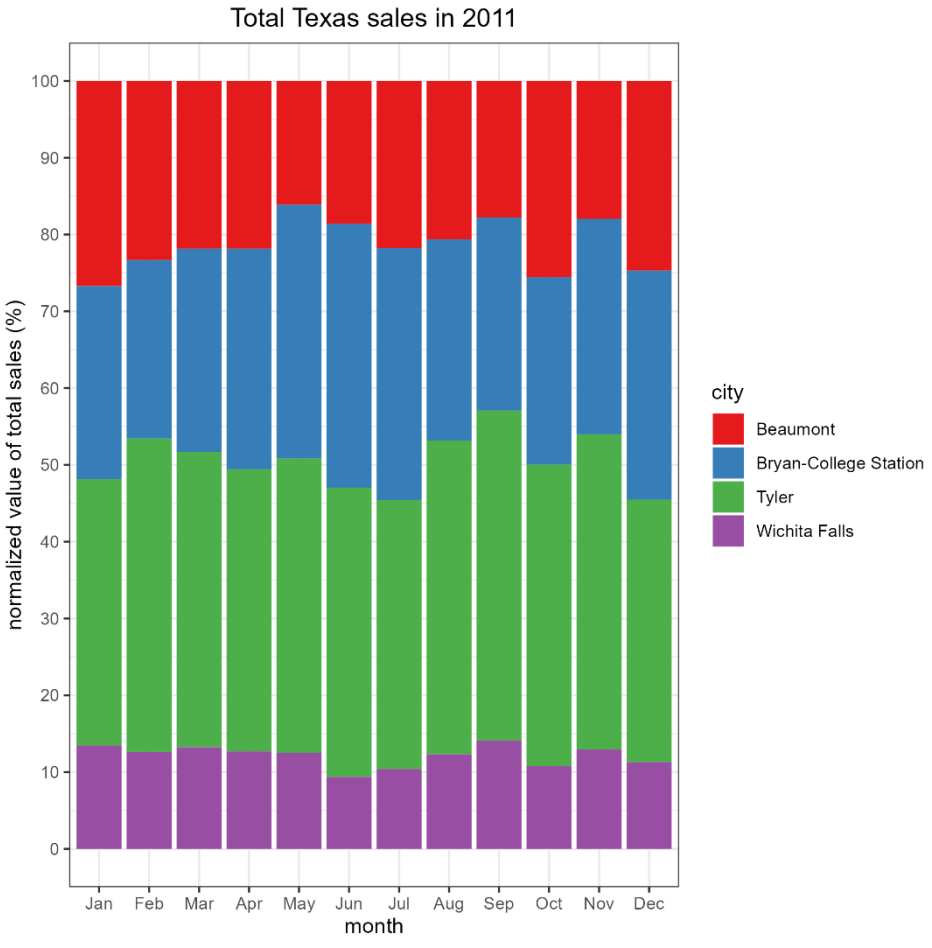


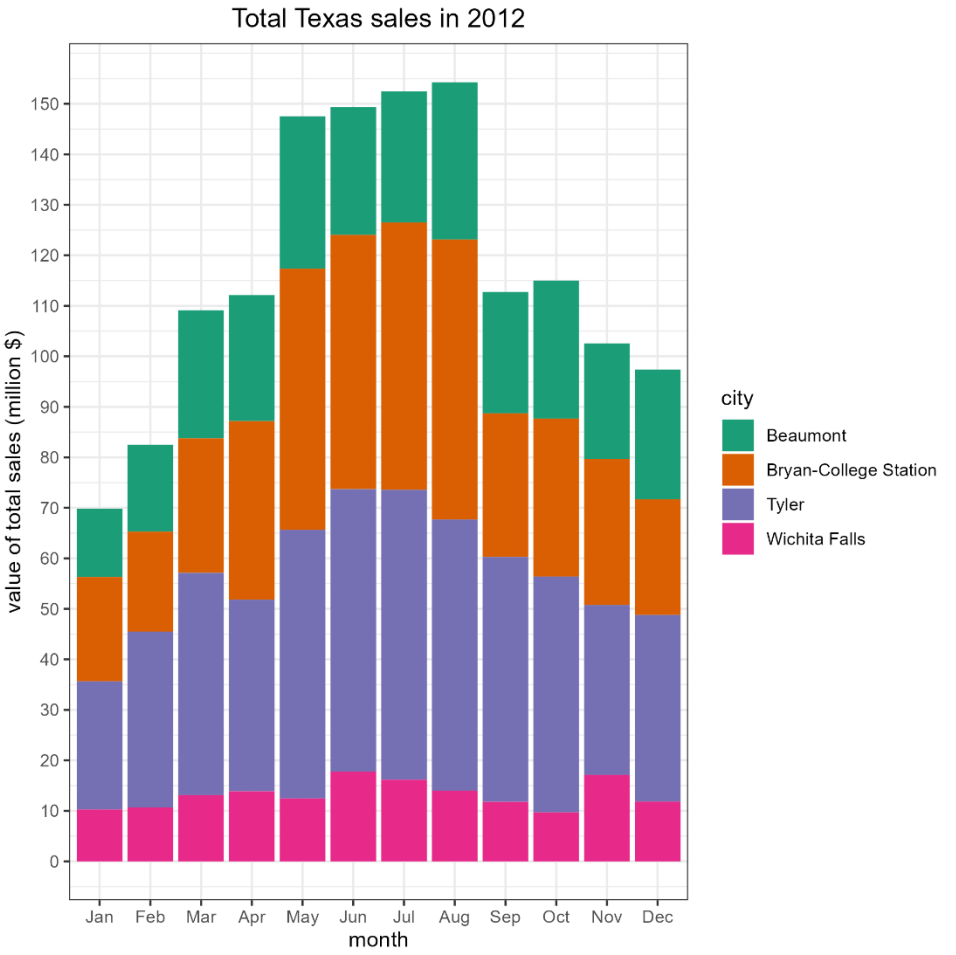
For what concerns the year 2011, the relationships among the cities that were described for 2010 remain unchanged. Anyway, the more profitable months for the Texas real estate market during this year are May, June, July, and August, with July replacing March (graph on the left).

The normalized comparison shows again how the proportions of "volume" changed only a little within the months of 2011:

1) Beaumont ~ 15-25 %;
2) Bryan-College Station ~ 20-35 %;
3) Tyler ~ 35-40 %;
4) Wichita Falls ~ 10-15 %;

Moreover, there is not a significant change in the percentages when compared to 2010 ones (graph on the right).



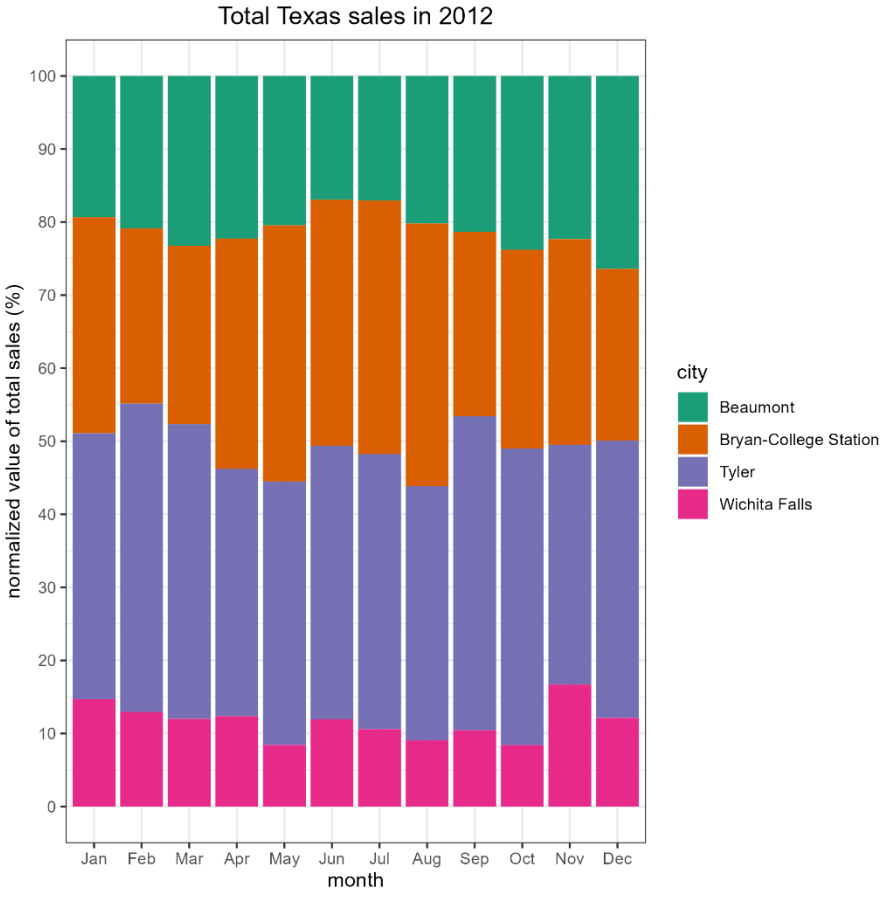


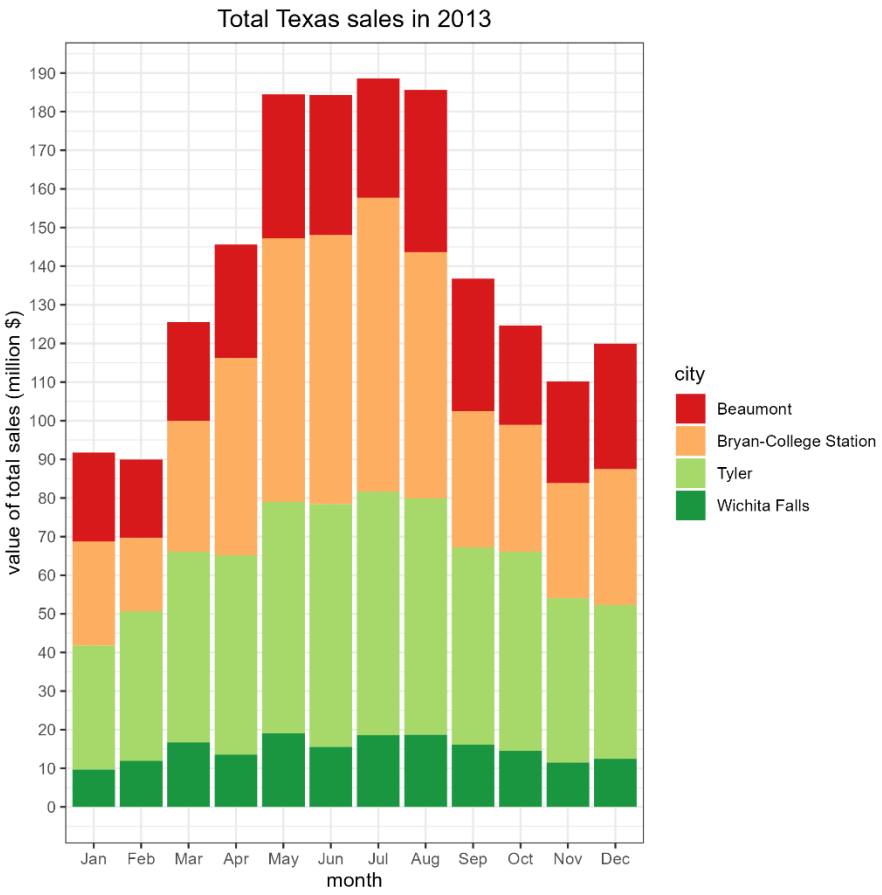
Also in 2012, the more profitable months for the Texas real estate market are May, June, July, and August. Important to notice how, during the current year the overall maximum value for the "volume" variable increases compared to the previous years (graph on the left).

Another time, the proportions characterizing the cities at the beginning of the year vary just a little with the passing of months.

1) Beaumont ~15-25 %;
2) Bryan-College Station ~ 25-35 %;
3) Tyler ~ 30-40 %;
4) Wichita Falls ~10-15 %;

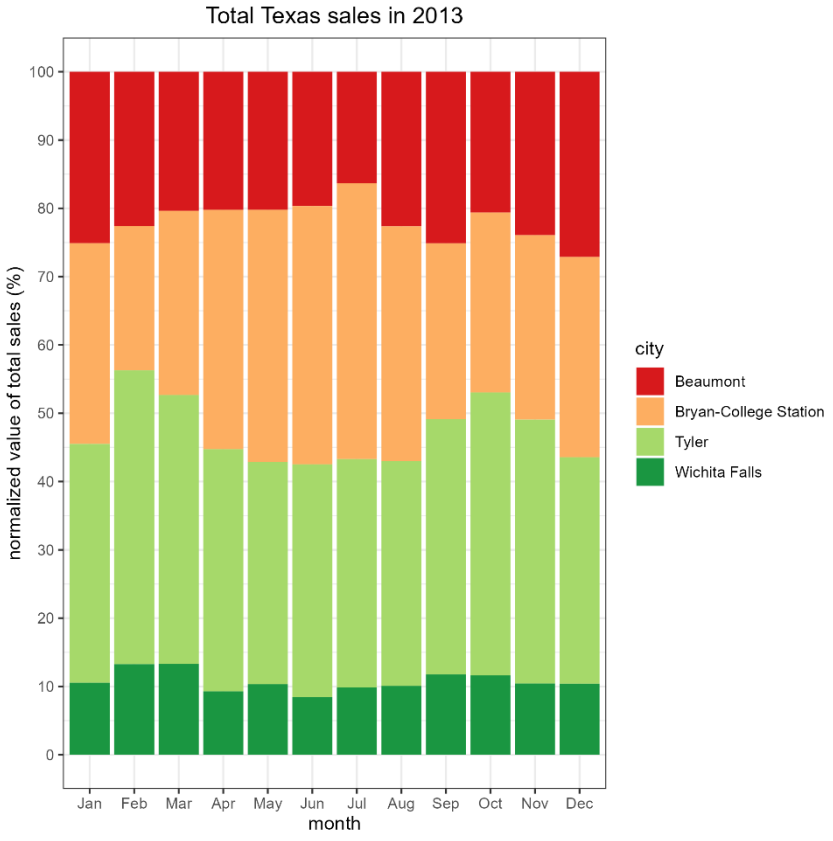
They also do not change much in respect to the past years.

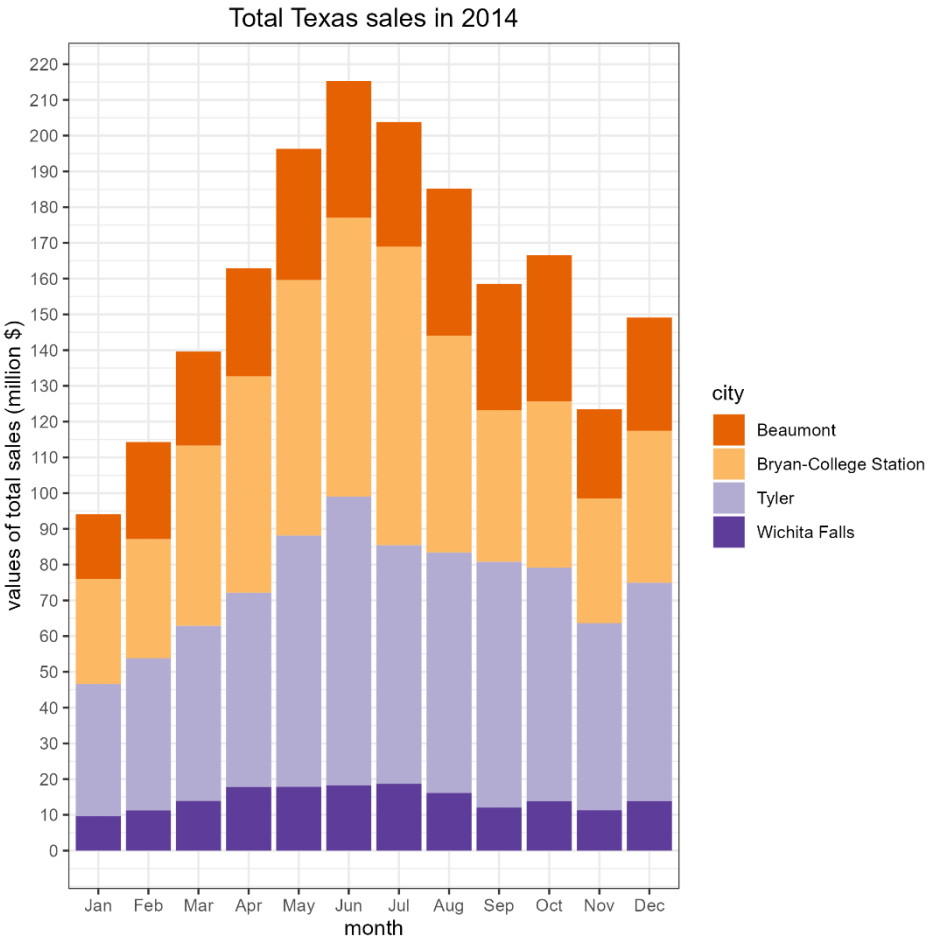




Even 2013 follows the same trend as the two previous years in terms of absolute and relative confrontation. There is a new global maximum value reached.

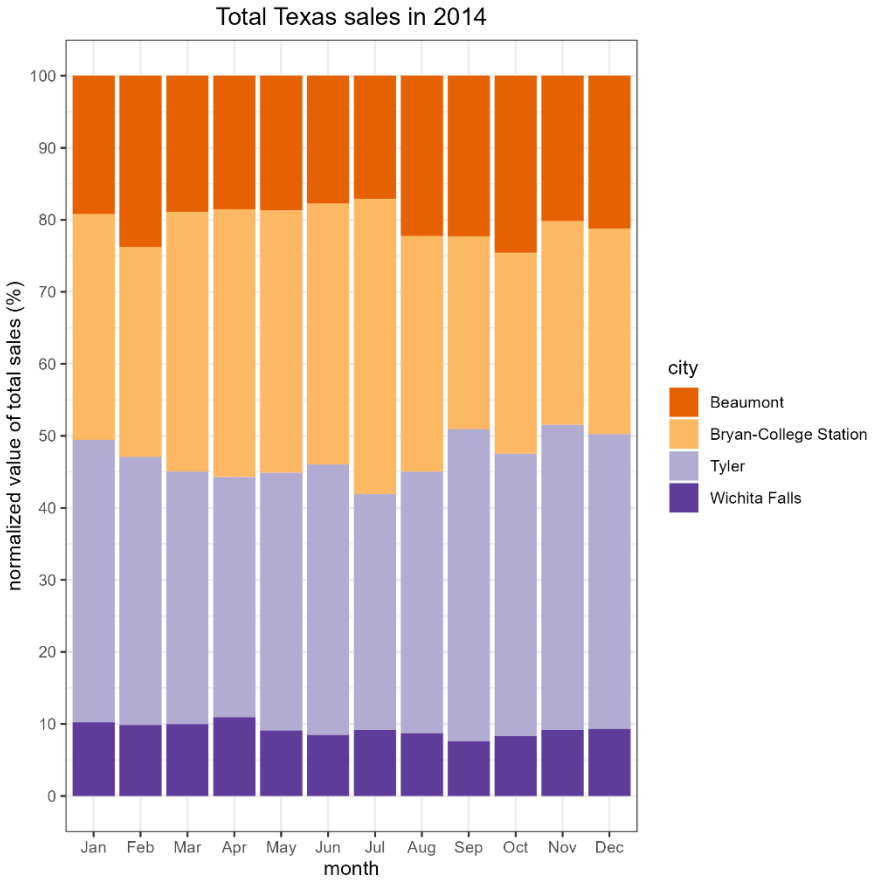
1) Beaumont
   ~15-25 %;
2) Bryan-College Station ~ 20-40 %;
3) Tyler ~ 35-45 %;
4) Wichita Falls
   ~10 %;





The aforementioned observations (2011-2013) are still valid for 2014. The overall peak values are further increased.

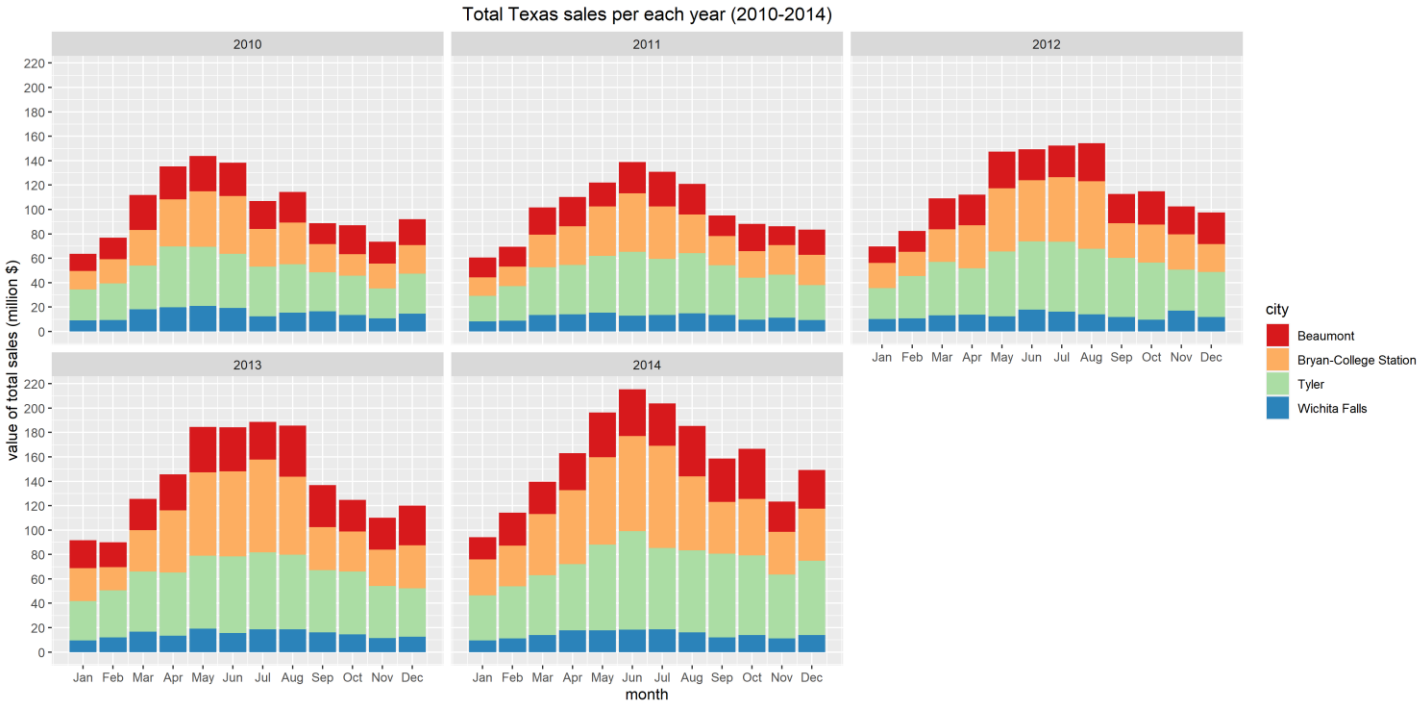
1) Beaumont ~20-25 %;
2) Bryan-College Station ~ 25-40 %;
3) Tyler ~ 35-40 %;
4) Wichita Falls ~10 %;



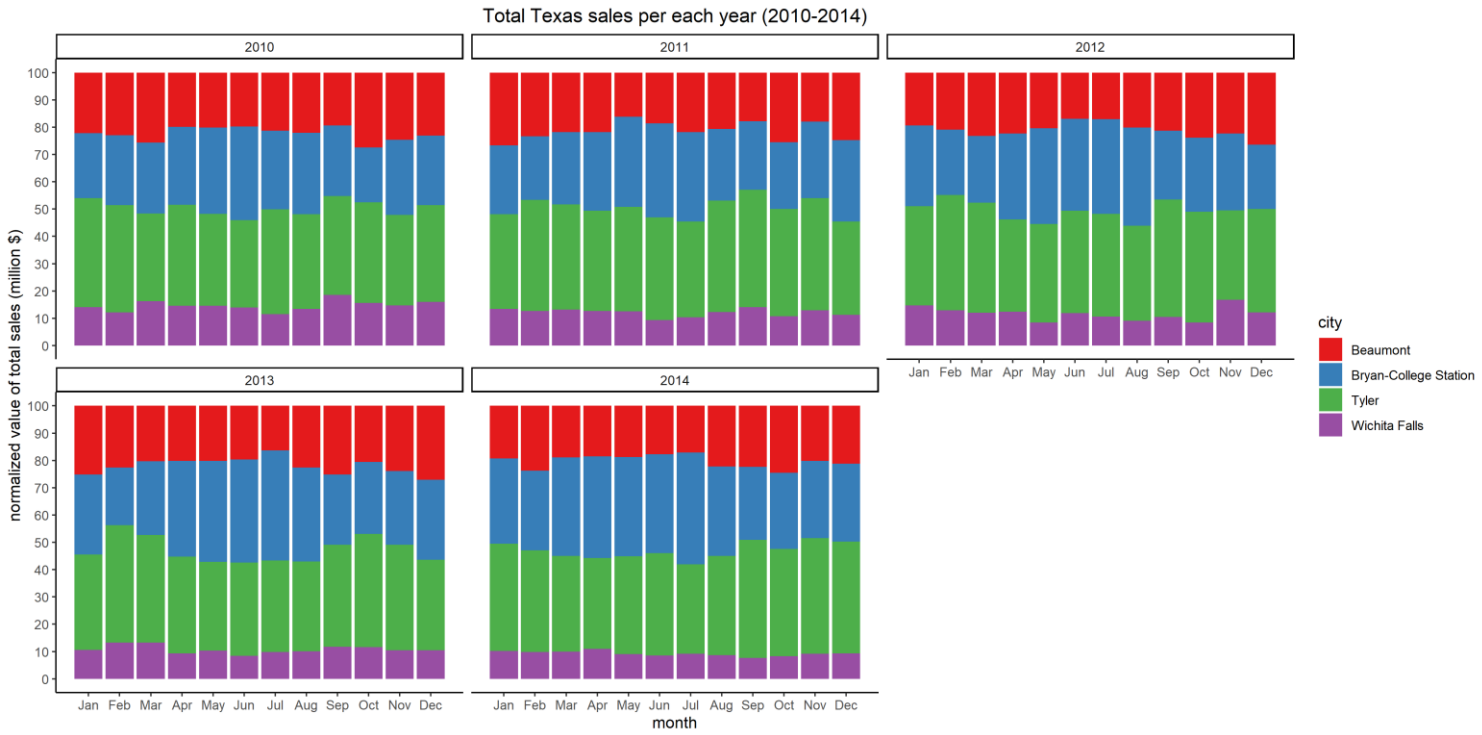
Display of the stack and normalized bar graph for the “volume” variable

To better visualize the data from the five years together also the “year” variable can be included in addition to “month” and ”city” in a geom_col graph. This was obtained through the *facet_wrap(~year)* command as previously shown for “list_efficacy”.

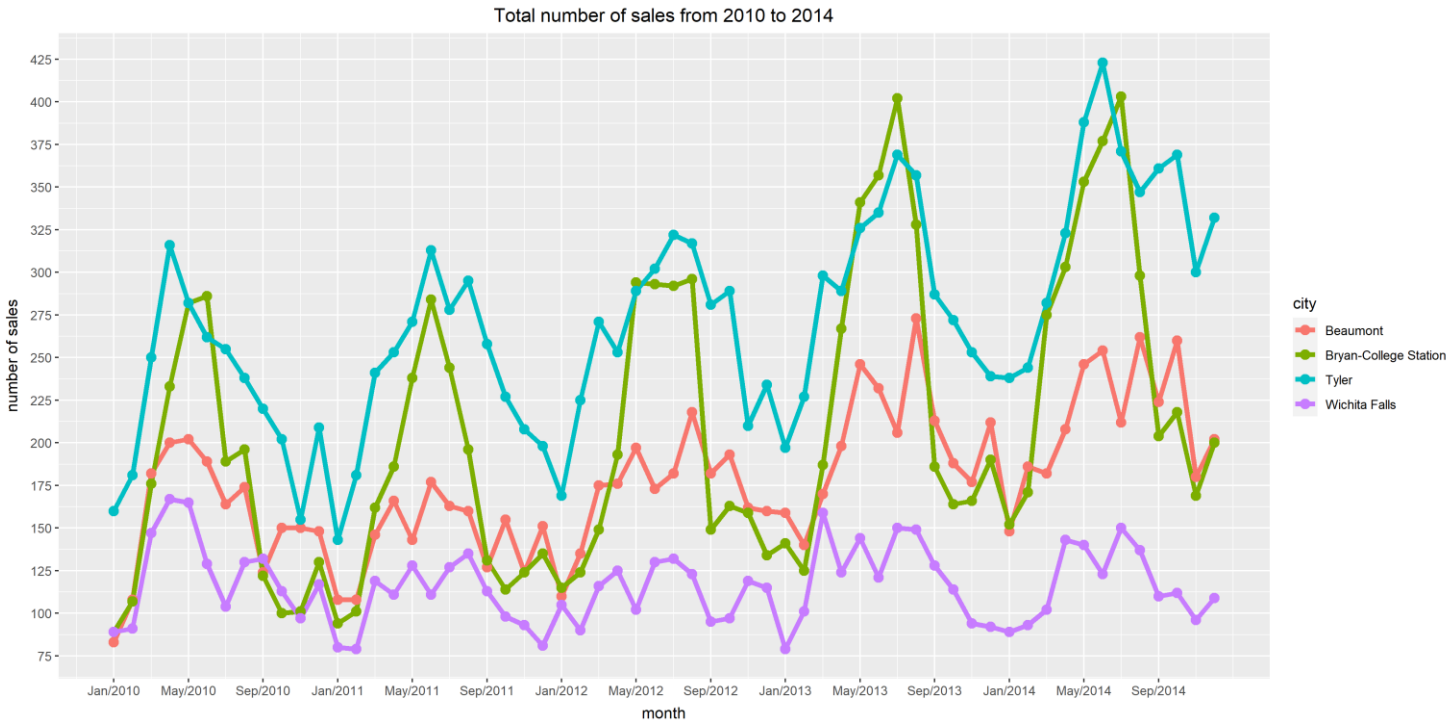
The stack bar graph better presents how the general increment of the maximum values that occurs after 2011 is due to the growth of the value of total sales in Tyler, Beaumont, and Bryan-College Station cities whereas Wichita Falls does not contribute much.



On the other side, the graph below better depicts how the relative proportions of “volume” in each month are interested by little changes over the five years.



Total number of sales from 2010 to 2014



In compliance with the trends characterizing the "volume" variable, the number of sales pursues the same pattern. There is a general increase during Winter and early Spring with the peak value in Spring-Summer followed by a decrement occurring within the subsequent months that stops around January, closing the cycle. Wichita Falls represents the most constant pattern that repeats itself pretty much over the years and together with Beaumont city, they are characterized by continuous fluctuations. On the contrary, Tyler and Bryan-College Station are interested by an increment in their maximum and minimum peak every year, despite following an annual cycle. Furthermore, the trend of Beaumont's number of sales decreases between 2010 and 2011, followed by its growth until 2013 and a subsequent stationary period.

Generally, it appears that "sales" changes together with "volume", explaining why the summer enlargement in the total value of the sales is based on a respective growth in the number of sold announcements.