Analysis of Real Estate in Texas

Luca Tripodi

September 2023

1 Dataset Introduction

This dataset provides an overview of real estate sales between 2010 and 2014 for some cities in Texas. It includes information such as city, month, number of listings and sales, which allow to analyze the trends for different variables and in various time-frames.

2 Variables Types

Below the types description for each variable:

- **"city":** categorical variable on nominal scale, denoting a Texan city.
- **"year":** categorical variable on ordinal scale, denoting the reference year.
- **"month":** categorical variable on ordinal scale, denoting the reference month. Since month is on circular scale, an eventual global ordering should consider also the reference year.
- **"sales":** numeric discrete variable, denoting the total number of sales for related city, year and month.
- **"volume":** numeric continuous variable, denoting the total value of sales expressed in millions of dollars [M$].
- **"median_price":** numeric continuous variable, denoting median price expressed in dollars [$].

- **"listings":** numeric discrete variable, denoting total number of active sale advertisements.
- **"months_inventory":** numeric continuous variable, denoting the amount of time expressed in months required to sell all current listings at current sales rate.

3 Measures of Central Tendency and Dispersion

For information on calculations please refer to the attached R script, below some consideration for each variable:

- **"city":** since it is a categorical variable on nominal scale, only the mode can be measured. Checking the frequency distribution, in total 4 modalities (4 different cities) are observed, each one with same cardinality equal to 60, so it's not present a single absolute mode. Moreover the distribution is characterized by the maximum heterogeneity.
- **"year":** sames observation as *city*, but in total 5 modalities (5 different years) each one with same cardinality equal to 48.
- **"month":** sames observation as *city* and *year*, but in total 12 modalities (all months of the year) each one with same cardinality equal to 20.
- **"sales":** all commons position and variability indexes have been calculated, see the code for details.
- **"volume":** all commons position and variability indexes have been calculated, see the code for details. It is worth mentioning that three modal values have been observed, but since their cardinality is equal to 2 (all other values appear only once), the distribution is almost at maximum heterogeneity.
- **"median_price":** all commons position and variability indexes have been calculated, see the code for details.
- **"listings":** all commons position and variability indexes have been calculated, see the code for details.
- **"months_inventory":** all commons position and variability indexes have been calculated, see the code for details.

Below some tables to summarize the data distributions for each variable, reporting the main descriptive indexes:

Position Indexes

	Min	1st Qu	Median	Mean	3rd Qu	Max
sales	79.0	127.0	175.5	192.3	247.0	423.0
volume	8.166	17.660	27.062	31.005	40.893	83.547
median_price	73800	117300	134500	132665	150050	180000
listings	743	1026	1618	1738	2056	3296
months_inventory	3.400	7.800	8.950	9.193	10.950	14.900

Dispersion Indexes

	Variance	Standard Deviation	Coefficient of Variation
sales	6344.3	79.65	41.42
volume	277.27	16.65	53.71
median_price	513572983	22662.15	17.08
listings	566569	752.71	43.31
months_inventory	5.31	2.3	25.06

Heterogenity and Shape Indexes

	Normalized Gini Index	Fisher Index	Kurtosis Index
sales	1	0.72	2.69
volume	1	0.88	3.18
median_price	1	0.36	2.38
listings	1	0.65	2.21
months_inventory	0.99	0.04	2.83

Counts for Categorical Variables

	Beaumont	Bryan-College Station	Tyler	Wichita Falls
city	60	60	60	60

	2010	2011	2012	2013	2014
year	48	48	48	48	48

	1	2	3	4	5	6	7	8	9	10	11	12
month	20	20	20	20	20	20	20	20	20	20	20	20

4 Variability and Asymmetry

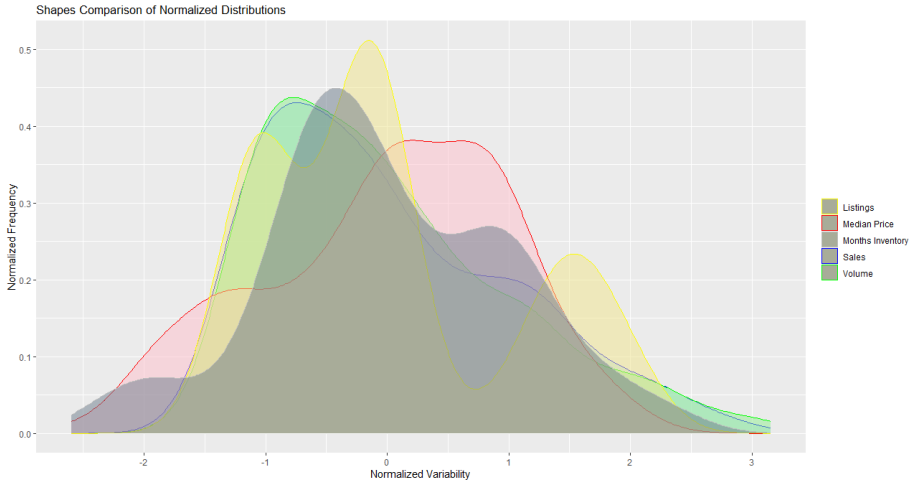
Since the domains of numeric variables consist only of positive values, by evaluating means and standard deviations for each variables it is meaningful to assess the coefficients of variation. Below their calculated values expressed in percentage:

```
sales_CV = 41.42
volume_CV = 53.71
median_price_CV = 17.08
listings_CV = 43.31
months_inventory_CV = 25.06
```

The *volume* appears to have the greatest variability among all the numeric variables.
To measure the asymmetry of a distribution the Fisher indexes can be evaluated, below the results:

```
sales_skew = 0.718
volume_skew = 0.885
median_price_skew = -0.364
listings_skew = 0.649
months_inventory_skew = 0.041
```

The *volume* variable exhibits the highest absolute value, indicating the greatest asymmetry among the variables. Furthermore, the signs of Fisher indexes were predicted by comparing the mean and median values: variables with means higher than their medians show positive asymmetry, so their distributions are skewed to the right; *median_price* variable is the only one characterized by a negative index and so with skewness to the left. By calculating the Z-scores so to standardize all the distributions, it becomes possible to compare their different shapes and consequently gain a better understanding of the considerations made regarding variability and skewness.



5 Class Division for Sales

Among the quantitative variables, the *sale* one has been divided in 5 classes. The discretization was performed using intervals of width 100, so defining frequency classes representing sales ranges. Taking advantage of the simplification into classes, the bar graphs related to absolute, relative, cumulative and cumulative relative frequencies have been displayed.

6 Gini Index

As previously mentioned when considering city heterogeneity, the Gini index for the *city* variable is equal to 1, as all four modalities in the dataset are all observed with the same frequency.

7 Probability

When randomly selecting a record from the dataset, there is a 25% probability of selecting an observation related to Beaumont city, an 8.33% probability of selecting one related to the month of July, and a 1.67% probability of selecting a record from December 2012.

8 Mean Price

A new column has been added to the dataset to represent the mean price, calculated by dividing the volume (converted into dollars) by the number of sales, both related to the same observation.

9 Listings Efficiency

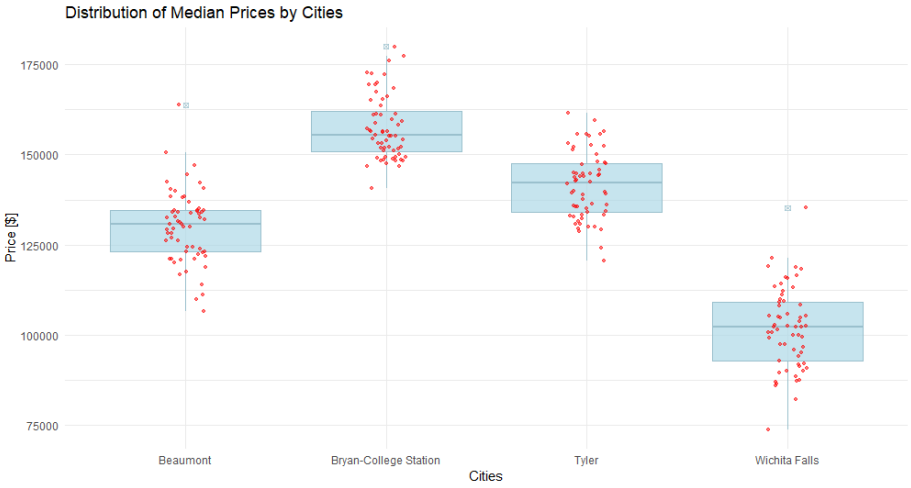
The efficiency of a listing can be measured by evaluating the ratio of the number of sales to the number of listings. This means that a listing can be considered more effective when it generates a higher number of sales, even when compared to another record with a similar number of listings.

10 Summaries and Plots

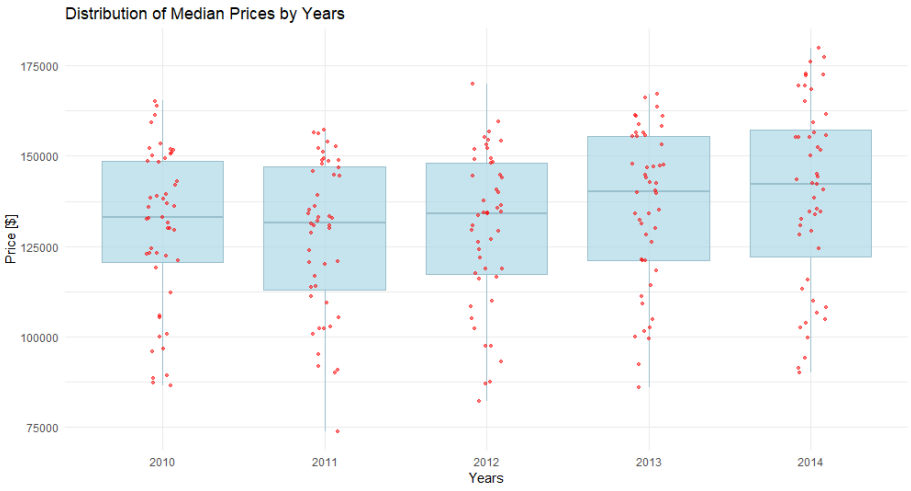
See the code for summaries about mean and standard deviation for sales, volume and listings, obtained by grouping separately by city, year and month.

10.1 Distributions of Median Prices by Grouping

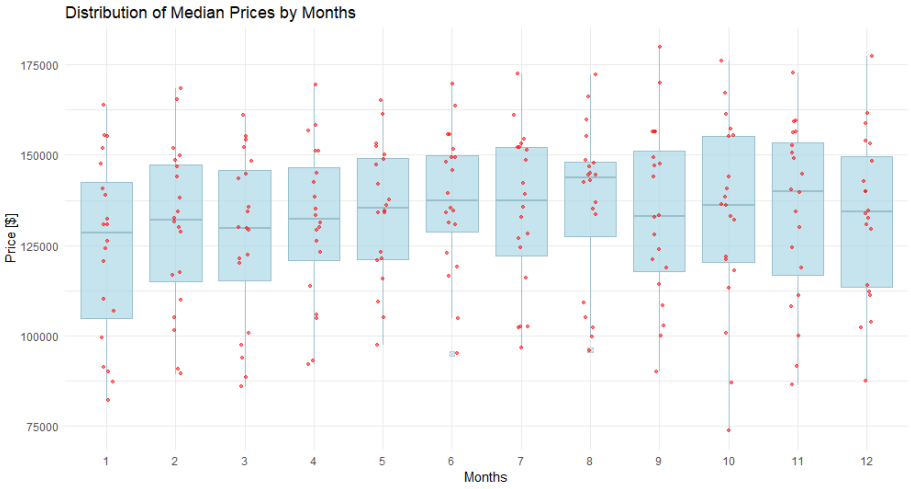
When visualizing the distribution of median prices for all cities using a boxplot representation, it can be observed that median prices are situated in various price ranges. However they all show relatively similar dispersion.



Comparing the boxplot displaying the distribution by years, the median prices appear more scattered, reflecting the different position indices of prices observed across cities.

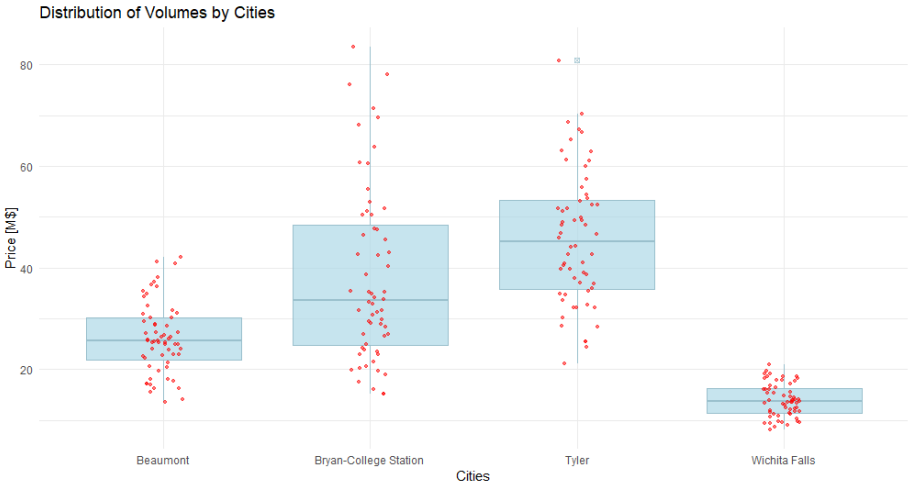


With finer division into months, the median prices appear to have a slightly increased variability, probably due to the reduced sample sizes for each month compared to the division by years.

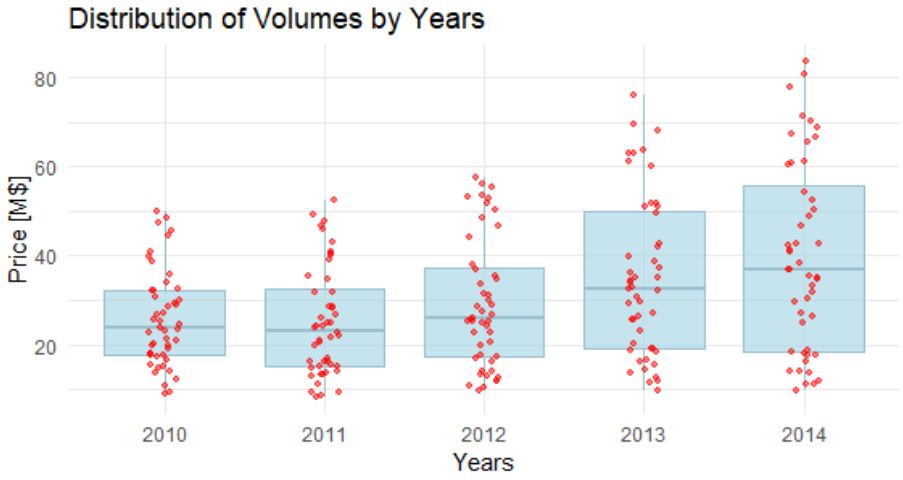


10.2 Distributions of Volumes by Grouping

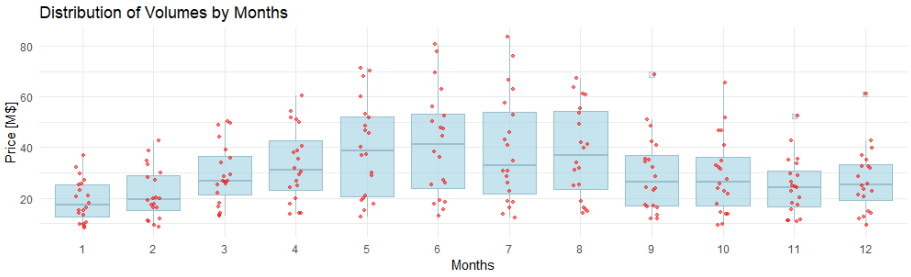
Comparing volume distributions across cities, it is observed a reduced variability in Beaumont and even narrower variability in Wichita Falls.



Volume distributions across the years seem to slightly increase in terms of dispersion as the years pass

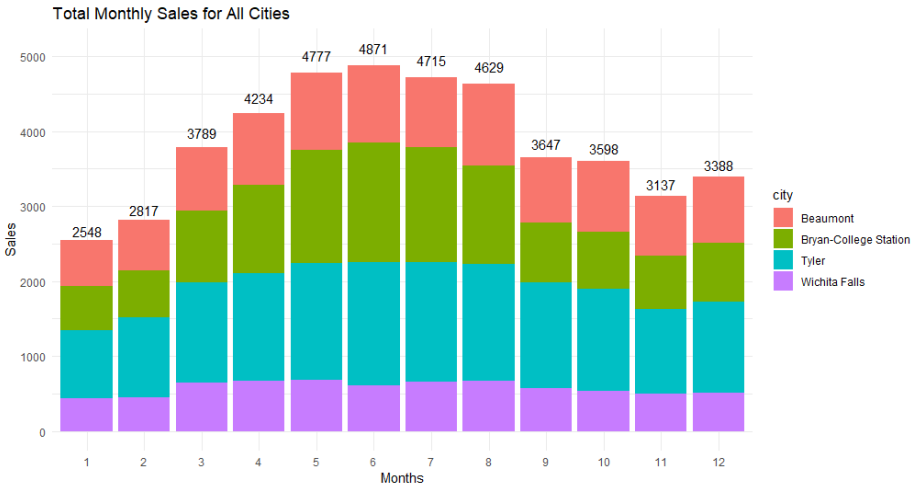


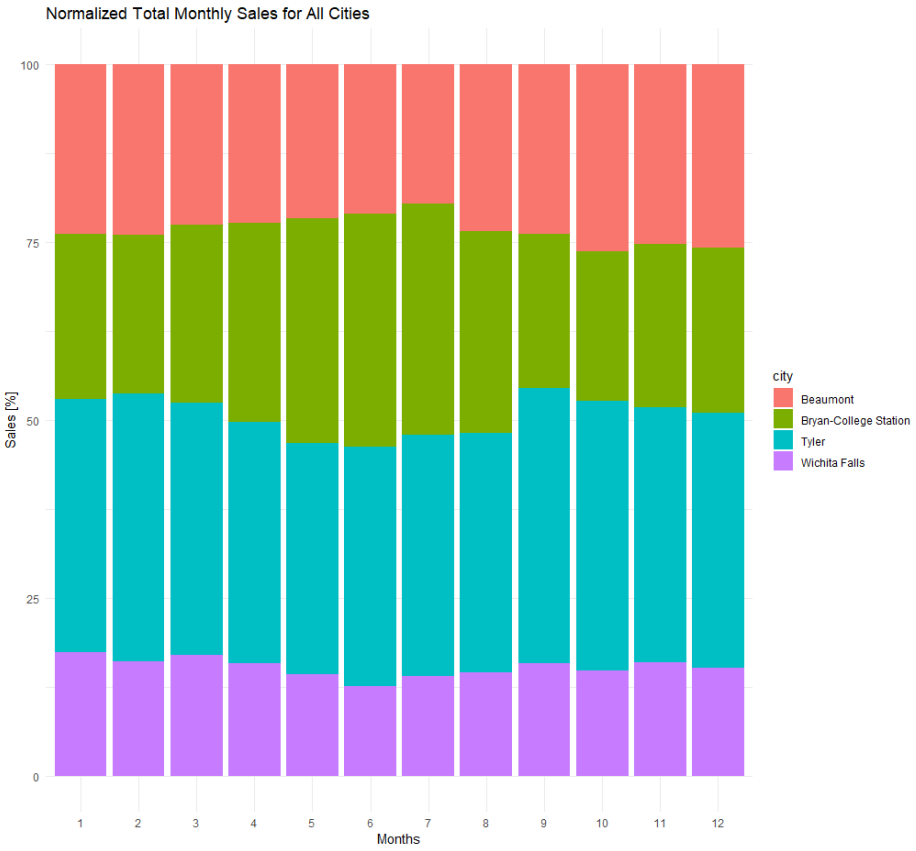
Volume distributions across the months seem to have a higher dispersion when the median price is higher



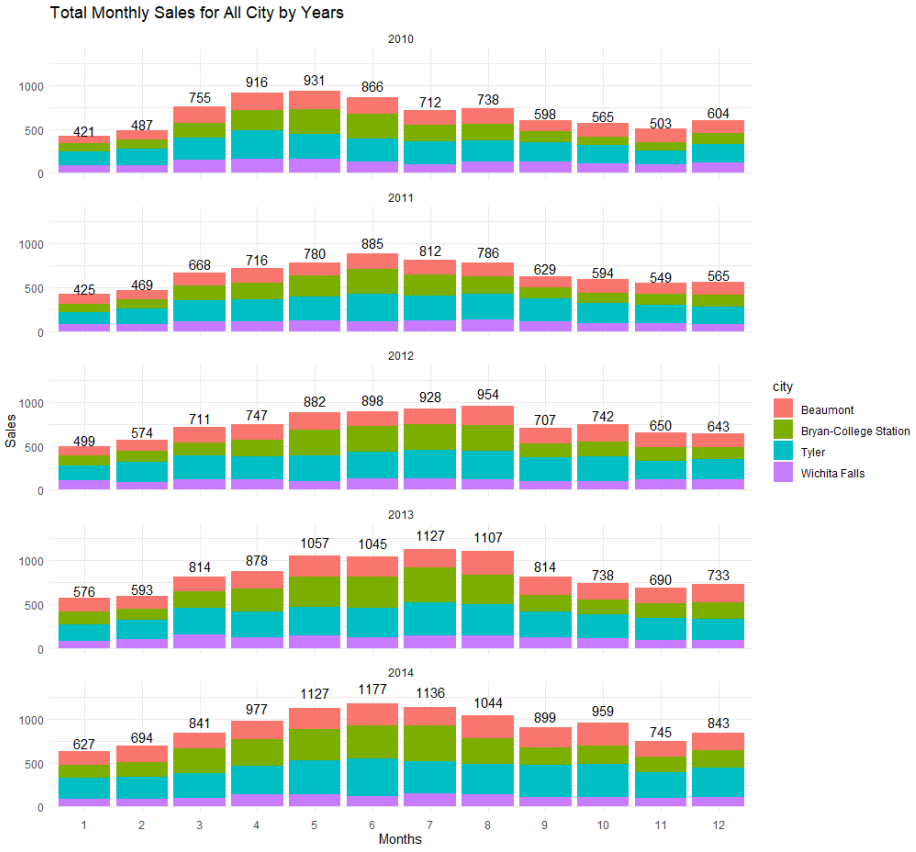
10.3 Total Monthly Sales for All Cities

Using a stack bar plot, the total sales for each month for all cities can be displayed. The graph shows a clear upward trend in the number of sales, starting from the beginning of the year and reaching its peak in June. Subsequently, there is a gradual decline until November. This pattern is particularly noticeable in Bryan-College Station and Tyler, the two cities with the highest absolute number of sales compared to the others.


Total Monthly Sales for All Cities
Sales
Months
2548
2817
3789
4234
4777
4871
4715
4629
3647
3598
3137
3388
city
Beaumont
Bryan-College Station
Tyler
Wichita Falls


Normalized Total Monthly Sales for All Cities
Sales [%]
Months
city
Beaumont
Bryan-College Station
Tyler
Wichita Falls

Splitting instead by year, quite the same trend curves are observed, but the maximum sales peak seems falling in a different month each year.



10.4 Time Series for Listings by Cities

Exploring the trends in monthly listings over the years, it becomes evident that these trends follow an annual periodicity, similar to what has been observed in sales trends. There is an upward trend from the beginning of the year until mid-year, followed by a decline towards year-end. This pattern is particularly pronounced in the city of Tyler, which has a significantly higher number of advertisements compared to the others. Additionally, it seems present a subtle decrease in the mean listings values over the years.