# Exploratory Analysis of the Texas Real Estate Market

Enrico Michelon

October 2023

## Contents



# 1 Introduction

This project concerns an exploratory analysis of the Texas real estate markets, in particular real estate in the period from 2010 to 2014 for the cities of Beaumont, Bryan-College Station, Tyler and Wichita Falls.

# 2 Dataset (Point 1,2,3,6)

The provided dataset consists of 240 objects of 8 variables which are *city, year, month, sales, volume, median price, listings, months inventory*. In this section will be discussed points 1, 2, 3 and 6 of the project. The main points are shown in bold.

## 2.1 City

**City is a qualitative variable on nominal scale** and represents the city where the property is for sale. Within the dataset there are 4 cities: Beaumont, Bryan-College Station, Tyler and Wichita Falls, which are equally distributed among the dataset, counting 60 rows each (Table 1). We can conclude that city has a four-mode distribution, since the absolute frequency is the same for all categories and therefore the **expected value of the Gini index is 1** (Point 6). Indeed:

$$G' = \frac{G}{(J-1)/J} = 1$$

where:

$$G = 1 - \sum_{j=1}^{J} f_j^2$$

| City | $n_i$ | $f_i$ | $N_i$ | $F_i$ |
|---|---|---|---|---|
| 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 |

**Table 1:** City distribution.

The R code used to obtain the results described above can be found in real_estate_market_analysis.R.

## 2.2 Year

**Year is a qualitative variable on ordinal scale**: represents the reference year for 2010 to 2014. For each city we can find 12 observations of each year. A preliminary consideration can be done observing the dataset: the year with the highest number of sales was the 2014, while the year with the lowest number of sales was 2011 (Table 2). Other considerations

| year | N° of sales |
|---|---|
| 2010 | 8096 |
| 2011 | 7878 |
| 2012 | 8935 |
| 2013 | 10172 |
| 2014 | 11069 |

**Table 2:** Sales per year.

can be found in Section 3.

## 2.3 Month

**Month is a qualitative variable (encoded) on ordinal scale**: represents the reference month for 1 (January) to 12 (December). The month of June (number 6) was in general the most prolific for sales, while January is the worst (Table 3).

| Month | Sales |
|---|---|
| 1 | 2548 |
| 2 | 2817 |
| 3 | 3789 |
| 4 | 4234 |
| 5 | 4777 |
| 6 | 4871 |
| 7 | 4715 |
| 8 | 4629 |
| 9 | 3647 |
| 10 | 3598 |
| 11 | 3137 |
| 12 | 3388 |

**Table 3:** Sales per month.

## 2.4 Sales

**Sales is a quantitative discrete** variable: represents the total number of sales per month. As we can observe from Table 4, the lowest number of sales registered in one month was 79, while the highest was 423. Looking at the first and third quartiles we obtained respectively 127 and 247. The mean value was 192.29 while the median value was 175.50.

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 79.00 | 127.00 | 175.50 | 192.29 | 247.00 | 423.00 |

**Table 4:** Position index for Sales.

Studying the variability of sales, we can see that the range of variation is $R = max(x) - min(x) = 344$, while considering just the central body, the interquariles range, computed as the difference between the third quartiles and the first, is 120. In particular, the dispersion of data around the mean is provided by variance and standard deviation, whose values are respectively: 6344.30 and 79.65. Finally the variability coefficient CV is 41.42.

| variance | std deviation | CV | IQR | range |
|---|---|---|---|---|
| 6344.30 | 79.65 | 41.42 | 120.00 | 344 |

**Table 5:** Variability measures for Sales.

A final consideration can be done observing the measures of shape, in particular skewness and kurtosis: **kurtosis value is -0.31 describing a *platykurtic* distribution**; skewness can be computed through the *Fisher skewness index*, whose value is 0.72, meaning that the **variable is positive asymmetric**. Indeed, the skewness value was expected to be greater than zero, since the mean is greater than the median. Further analysis corcerning "Sales" can be found in Section 3

## 2.5 Volume

**Volume is a quantitative variable on ratio scale**: represents the total value of sales in millions of dollars per month. Same considerations made for variable "Sales" can be made for variable "Volume". Position measures for volume variable can be found in Table 6, while measures of variability are given in Table 7

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 8.17 | 17.66 | 27.06 | 31.01 | 40.89 | 83.55 |

**Table 6:** Position measures for Volumes.

| variance | std deviation | CV | IQR | range |
|---|---|---|---|---|
| 277.27 | 16.65 | 53.71 | 23.23 | 75.38 |

**Table 7:** Variability measures for Volumes.

Finally shape measures report a skewness of 0.884742 meaning that the **variable is positive asymmetric** according to the trend of sales, and a **kurtosis of 0.176987, which means that the distribution is *leptokurtic***

## 2.6 Median Price

**Median price is a quantitative variable on ratio scale**. It represents the median sales price in dollars. Measures for position and variability are given in Table 8 and Table 9.

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 73800.00 | 117300.00 | 134500.00 | 132665.42 | 150050.00 | 180000.00 |

**Table 8:** Position measures for Median Price.

| variance | std deviation | CV | IQR | range |
|---|---|---|---|---|
| 513572983.09 | 22662.15 | 17.08 | 32750.00 | 106200.00 |

**Table 9:** Variability measures for Median Price.

Corcerning shape measures, **"Median Price" is a platykurtic variable**, since Kurtosis value is -0.6229618, while the Fisher-Pearson coefficient of **skewness is -0.3645529, meaning than this is a negatively skewed variable**, as expected since mean measure is lower than median measure. It is interesting to note that while the number of sales and volume have a tendency towards a positive asymmetry, in which most values are concentrated below the mean value, the median price has the opposite trend, in which there is a greater concentration of above-average values.

## 2.7 Listings

**Listings is a quantitative discrete variable**. It represents total the number of active listings. Measures for position and variability are given in Table 10 and Table 11.

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 743.00 | 1026.50 | 1618.50 | 1738.02 | 2056.00 | 3296.00 |

**Table 10:** Position measures for Listings.

| variance | std deviation | CV | IQR | range |
|---|---|---|---|---|
| 566568.97 | 752.71 | 43.31 | 1029.50 | 2553 |

**Table 11:** Variability measures for Listings.

From position measure we expect **"Listings" to be a positively skewed variable**. Indeed Fisher-Pearson skewness coefficient is 0.6494982. Kurtosis values is -0.79179, since this variable has a **platykurtic distribution**.

## 2.8 Month Inventory

**Month inventory is a quantitative on a scale of ranges**: represents the amount of time required to sell all current listings at the current rate of sales, expressed in months.

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|---|---|---|---|---|---|
| 3.40 | 7.80 | 8.95 | 9.19 | 10.95 | 14.90 |

**Table 12:** Position measures for Month Inventory.

| variance | std deviation | CV | IQR | range |
|---|---|---|---|---|
| 5.31 | 2.30 | 25.06 | 3.15 | 11.50 |

**Table 13:** Variability measures for Month Inventory.

Finally the variable has a **positively swekness and Fisher coefficient value is 0.04097527**, with a **platykurtic distribution, with a Kurtosis value of -0.1744475**.

## 2.9 Other considerations (Point 4)

In this section we give a description of the dataset, in terms of variables and position, variance and shape measures. We saw that sales, volumes, listings and month inventory have a positive asymmetric distribution, while median price has a negative asymmetric distribution. **Fisher-Pearson skewness coefficient indicates that sales has the most asymmetric distribution**. Kurtosis coefficient describes sales, listings, month inventory and median price as platykurtic distributions, while volume is described as a leptokurtic distribution. This means volume distribution presents longer tails, with an higher percentage of outlayers. Observing the coefficient of variability CV, we can observe that the median price is the variable has the lowest variability, while the **variable with the greatest variability is volume**. In Figure 1 we can see difference from positively skewed distribution for "Sales" and negatively skewed distribution for "Median price".

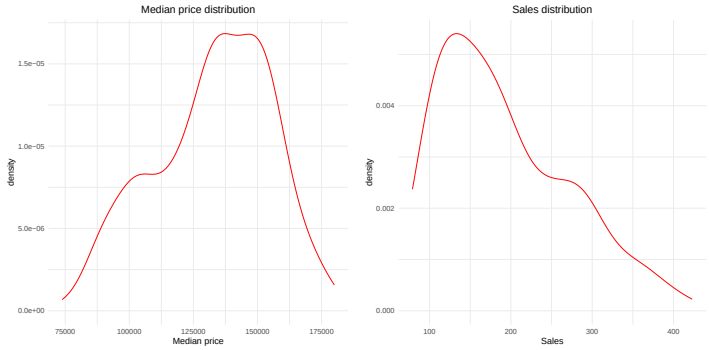


**Figure 1:** "Median price" and "Sales" probability function.

# 3 Analysis (Point 5)

To get a better idea of the pace of house sales, for Point 5 we have chosen to divide the variable "Month inventory" into classes, and obtained that most properties would sell in a period between 6 and 9 months (see Figure 2). The Gini coeffient can be easily computed as made in 2.1, and the value obtained is 0.85.

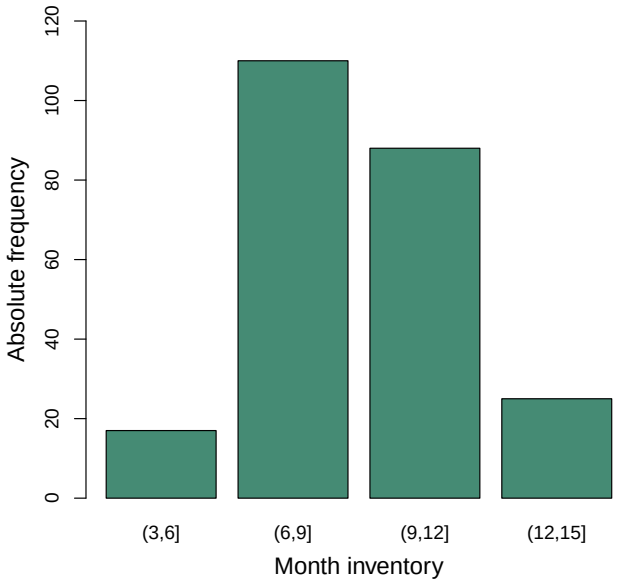


**Figure 2:** Frequency distribution of "Month inventory" class.

## 3.1 Average price and Effective ads (Points 8 and 9)

Exploiting the dataset, in particular variables "Volume" and "Sales" we added another column with the mean price information. The new dataset obtained can be found on the project directory in *data/real_estate_texas_with_mean_and_eff.csv*. In the same dataset a column with an effective advertisements rate can be found. It has been computed as a coefficient from 0 to 1, where 1 means the maximum effectiveness, 0 means minimum effectiveness, computed as "Sales" over the variable "Listings". This new variable provide a more unbiased instrument to compare sales between cities, or years or month. For example, in Figure 3 and Figure 4, we can observe the sales performance for the Texas cities under consideration during the period 2010-2014: if we only consider the number of sales, Tyler is the city with the highest number of sales. However, if we consider the coefficient of effectiveness of listings, we see that Tyler is the worst city in terms of the ratio of sales to listings, while the city with the best ratio is Bryan-College Station. This means that Tyler has the highest number of sales only because it is the city with the highest number of advertisements. Another consideration can be made for Wichita Falls: from the sales performance, it appears to be the

city with the lowest sales performance, yet the effectiveness of the advertisements has not deteriorated over time, and is superior to Beaumont and Tyler. Finally for Beaumont city trends of effectiveness and trend of sales have a very similar behaviour.

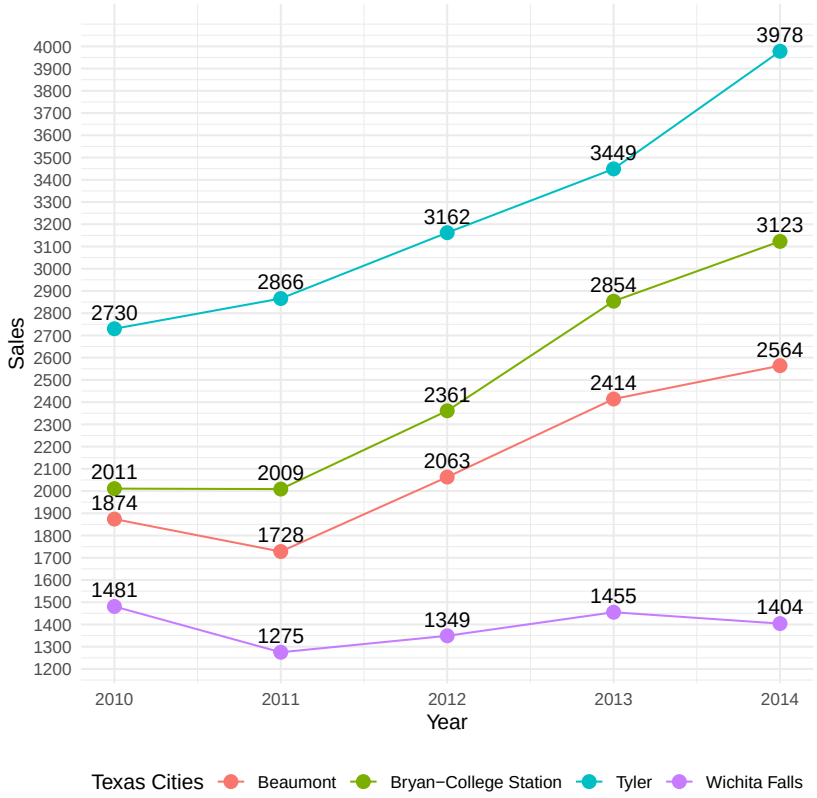


**Figure 3:** Sales trend.

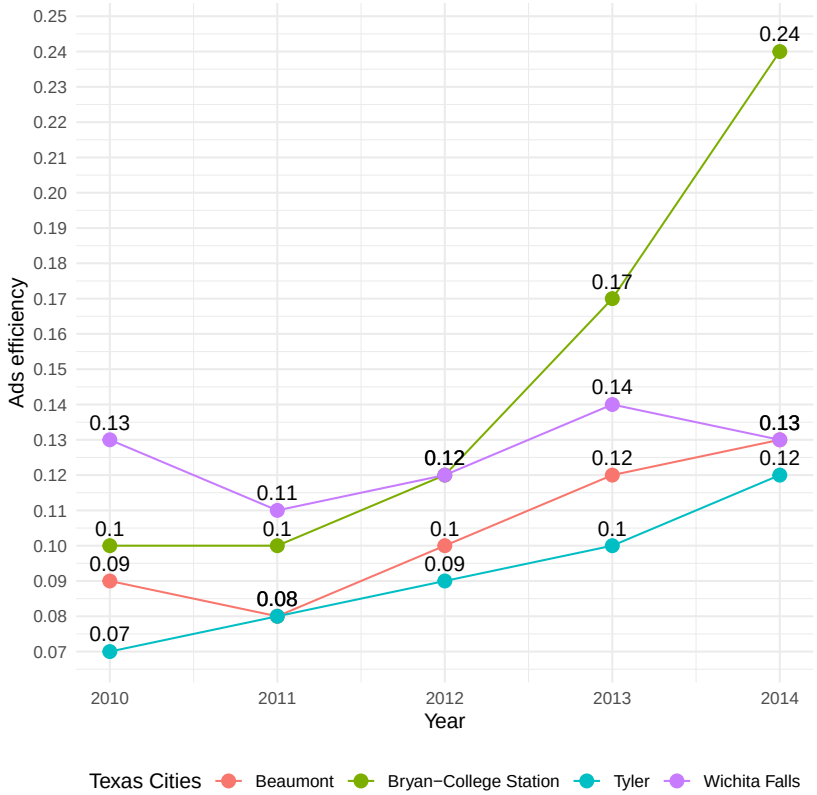


**Figure 4:** Ads effectiveness.

## 3.2 Probability (Point 7)

As required by the project, some probabilities were calculated. The probability of Beaumont being drawn at random, choosing one row of the dataframe, is 0.25, while the probability of July being drawn is 0.08. Finally, the probability of December 2012 being drawn is 0.017.

## 3.3 Summary (Point 10)

At this point we can observe some other phenomena, for example "Sales" trend, coditional on years and months (others can be found in the code). In Table 14, we can observe position measures and standad deviation for variable "Sales" conditional on years and months.

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Std Deviation |
|---|---|---|---|---|---|---|
| 421.00 | 621.25 | 743.50 | 769.17 | 898.25 | 1177.00 | 192.18 |

**Table 14:** General position measures for Sales with Standard deviation.

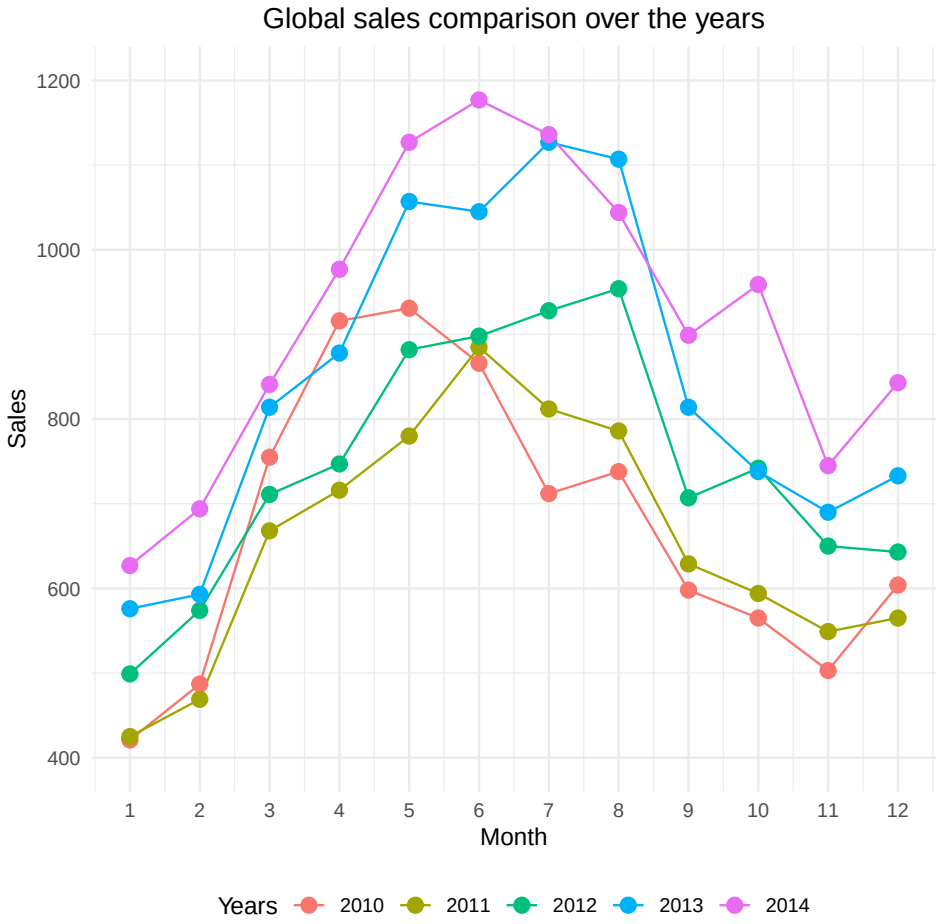


**Figure 5:** Sales comparison over the years.

From Figure 5 we can observe that the peak of sales has been registered in June 2014, while the lowest number of sales was on January 2010. Another interesting analysis can be made for variable "Volume". In Table 15 we can see position meausers for "Volume", grouped by city and year.

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Std Dev |
|---|---|---|---|---|---|---|
| 144.62 | 234.77 | 355.44 | 372.06 | 479.01 | 715.22 | 170.28 |

**Table 15:** Position measures for Volumes, grouped by year and city.

In Figures 6, 7, 8 and 9 we can see, for each year from 2010 to 2014, the distribution of

volumes per city, the average price of each property over the same period, the number of sales and the effectiveness of the advertisements. We can observe that even if Tyler has the lowest advertisement effectiveness, the city has the highest number of sales (due to the highest number of listings), the second highest mean price, and makes the largest contribution in terms of total sales value in millions of dollars.

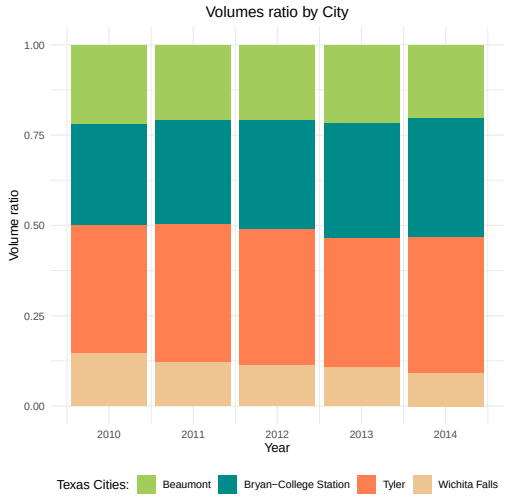


**Figure 6:** Volumes ratio by City.

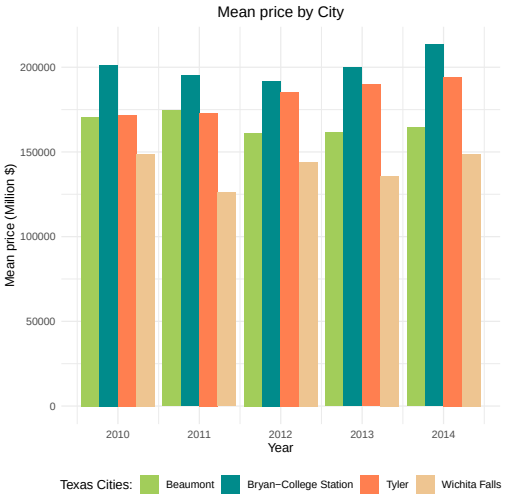


**Figure 7:** Mean price by City.

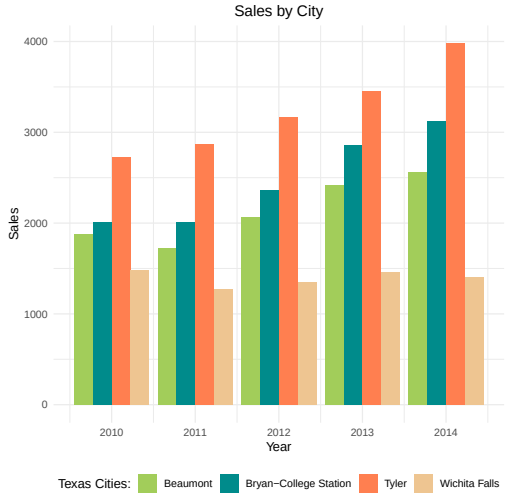


**Figure 8:** Sales by City.


**Figure 9:** Ads effectiveness by City.

Another interesting point is that Wichita Falls maintains a constant trend concerning efficacy, number of sales and mean price of properties, but makes an ever decreasing contribution from year to year in terms of total sales value in millions of dollars. Finally we can do some considerations for Bryan-College station. Table 16 shows the position measures and standard deviation, considering the average volumes of each month, in the years 2010-2014.

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. | Std Deviation |
|---|---|---|---|---|---|---|
| 21.49 | 29.24 | 31.99 | 38.19 | 50.69 | 58.64 | 13.86 |

**Table 16:** Position measures and standard deviation for Bryan-College Station (by month).

In Figure 10 average volumes of each month, in the years 2010-2014 are represented.

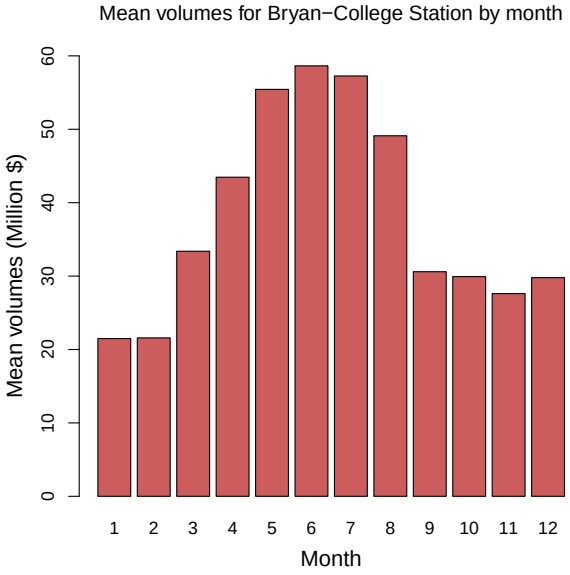


**Figure 10:** Mean volumes for Bryan-College Station by month.

# 4 More Graphics (Part 2)

In the second part, we are going to:

- use boxplots to compare the median house price distribution between cities;
- use boxplots to compare the distribution of the total value of sales between cities and years;
- compare total sales between cities over the months;
- create a line chart of volumes by city over the months, considering the entire period from 2010 to 2014.

## 4.1 Boxplots (Point 1, 2)

In **Figure 11**, we can observe the median house price distribution between cities. In particular, we can see that Bryan-College Station has the distribution with the highest values, while Wichita Falls has the distribution with the lowest values. Moreover, Tyler seem to have no outlayers, while Beaumont and Wichita Falls has some outlayers very distant from the rest of the distribution, with values 163800 and 135300 dollars respectively. Another consideration can be made in terms of interquartile range, since Wichita Falls has an interquartile range of 16375 dollars, while Bryan-College station has an interquartile range of 11175. Boxplot also provides an idea of the skewnees: for example Beaumont has a median price negatively distributed. In **Figure 12** instead, we can observe the boxplots that compare the distribution of the total value of sales between cities and years. As we can see, there is no outlayers in sales distribution. Moreover the distribution for Bryan-College Station is very scattered or sparse, specially in 2012 and 2013. On the other hand, Wichita Falls presents a distribution with a constant behavior over the years, and values which are more clustered. Another interesting point to note is the median value of sales for Bryan-College Station in 2012 and 2013, which is respectively 161 and 188, close to first interquartile (145 and 166 respectively), and distant to third interquartile (292, 331 respectively), representing a very strong positive skewness.

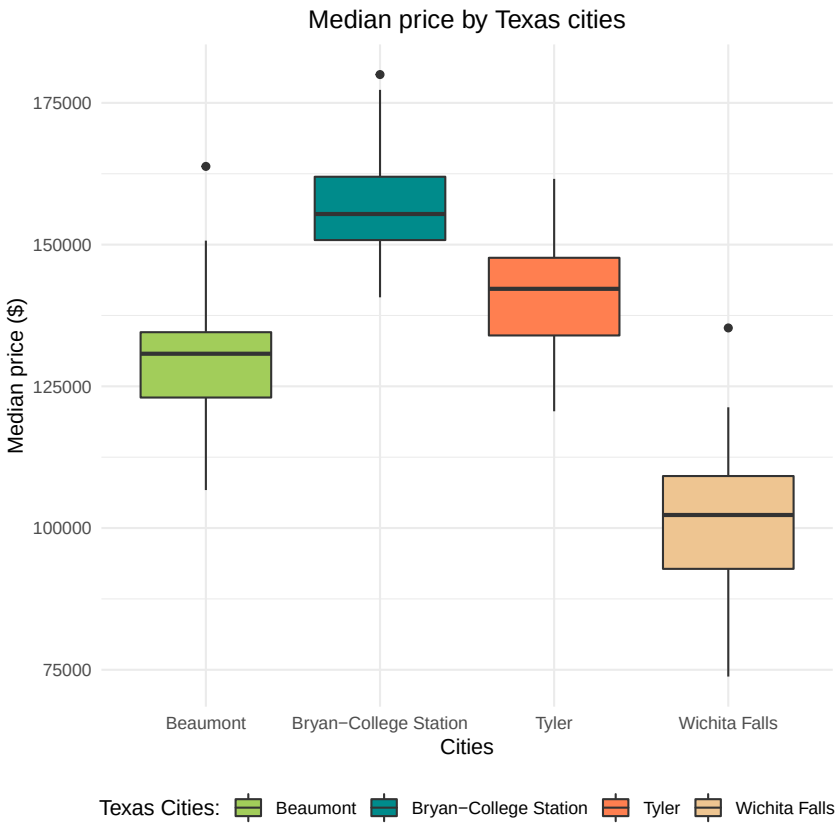


**Figure 11:** Median price by Texas cities.

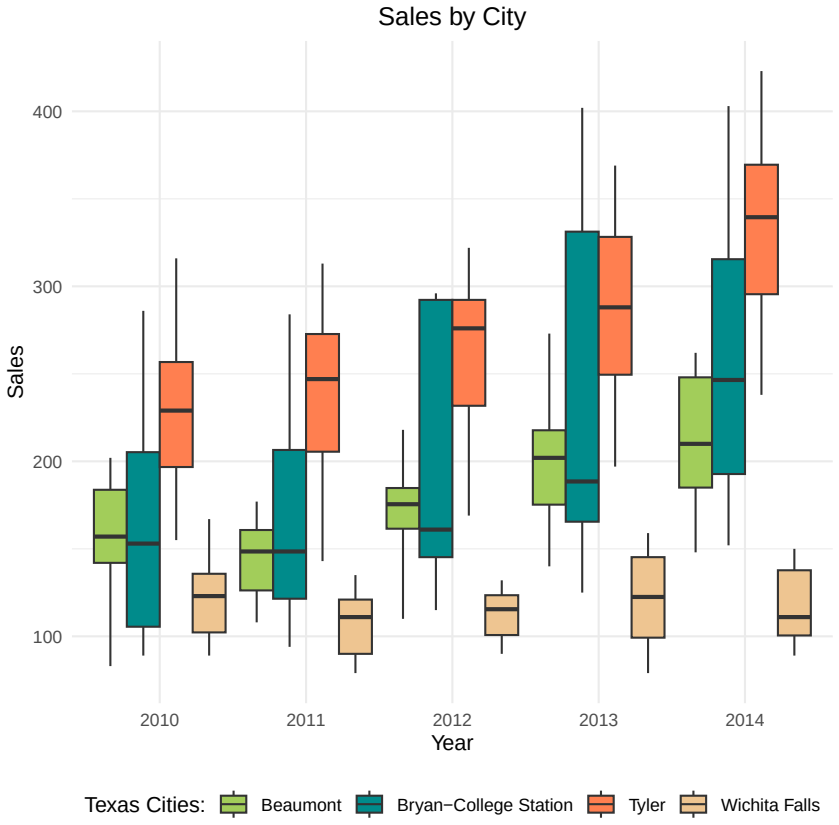


**Figure 12:** Sales comparison over the years.

To improve our observations, we can compare the graph of the probability function of Bryan-College Station and Wichita Falls in Figure 13. The graphs effectively show a positively skweness and a right tail for Bryan-College Station, while for Wichita Falls show a platykurtic distribution, with no skewness.

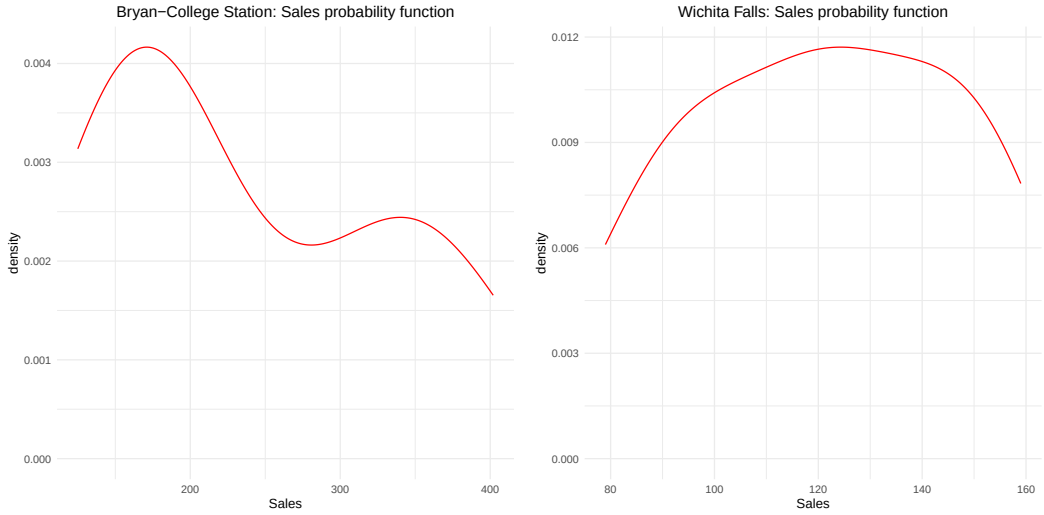


**Figure 13:** Sales probability function difference between Bryan-College Station and Wichita Falls (2013).

### 4.2 Stacked bar graph (Point 3)

Now we want to compare the total sales between cities over months. In Figure 14 we can see a normalised stacked bar graph, representing the ratio of sales per city over months, for each year from 2010 to 2014. We can see that Wichita Falls makes a constant sales contribution over the months and years. Bryan-College station, on the other hand, improves its contribution in the middle months (May, June and July), while making a decreasing contribution at the beginning and end of the year. In December, Beaumont contributes about 25% of annual sales.
Finally, Tyler is generally the city that makes the largest contribution.

### 4.3 Line chart (Point 4)

In Figure 15 sales trend for each city during the period from 2010 to 2014. Once again we can observe the discontinuity of sales for Bryan-College Station and Tyler, with a positive spike in sales in the middle months of the year, while the behaviour for Wichita Falls and Beaumont is more constant. Moreover both Bryan-College Station and Tyler show an increasing trend for sales.

## 5 Conclusion

We have performed an exploratory analysis of the Texas real estate markets, in particular real estate, in the period from 2010 to 2014 for the cities of Beaumont, Bryan-College Station, Tyler and Wichita Falls. We have seen different behaviour in "Sales", "Volume", "Median Price" and "Month Inventory" and "Listings" over months and years. We performed analysis also for "Mean Price" and "Effectiveness" of advertisements, combining measures of position, varibility, and shape to graphs.

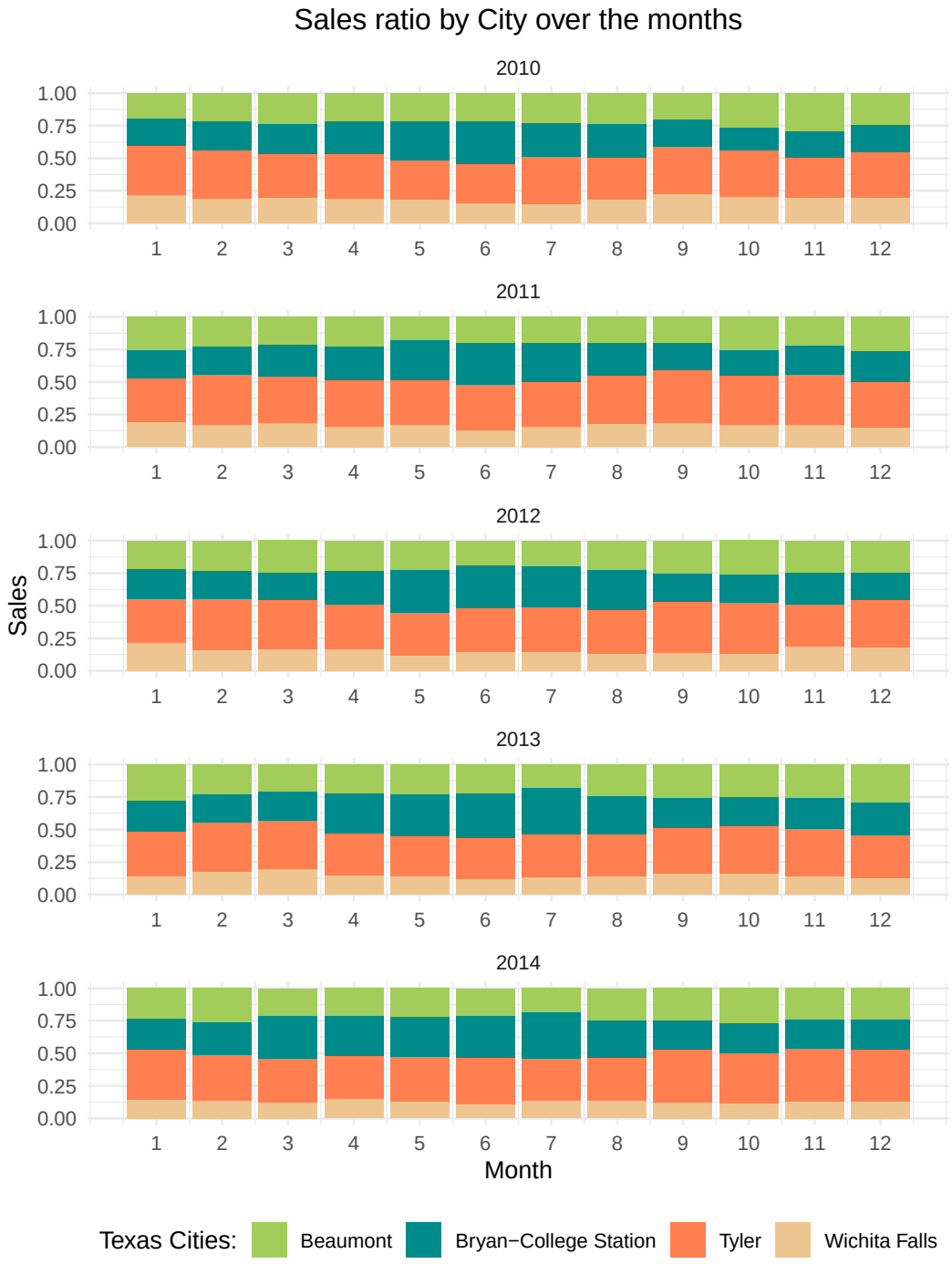


**Figure 14:** Sales ratio by cities over the months for 2010 to 2014.

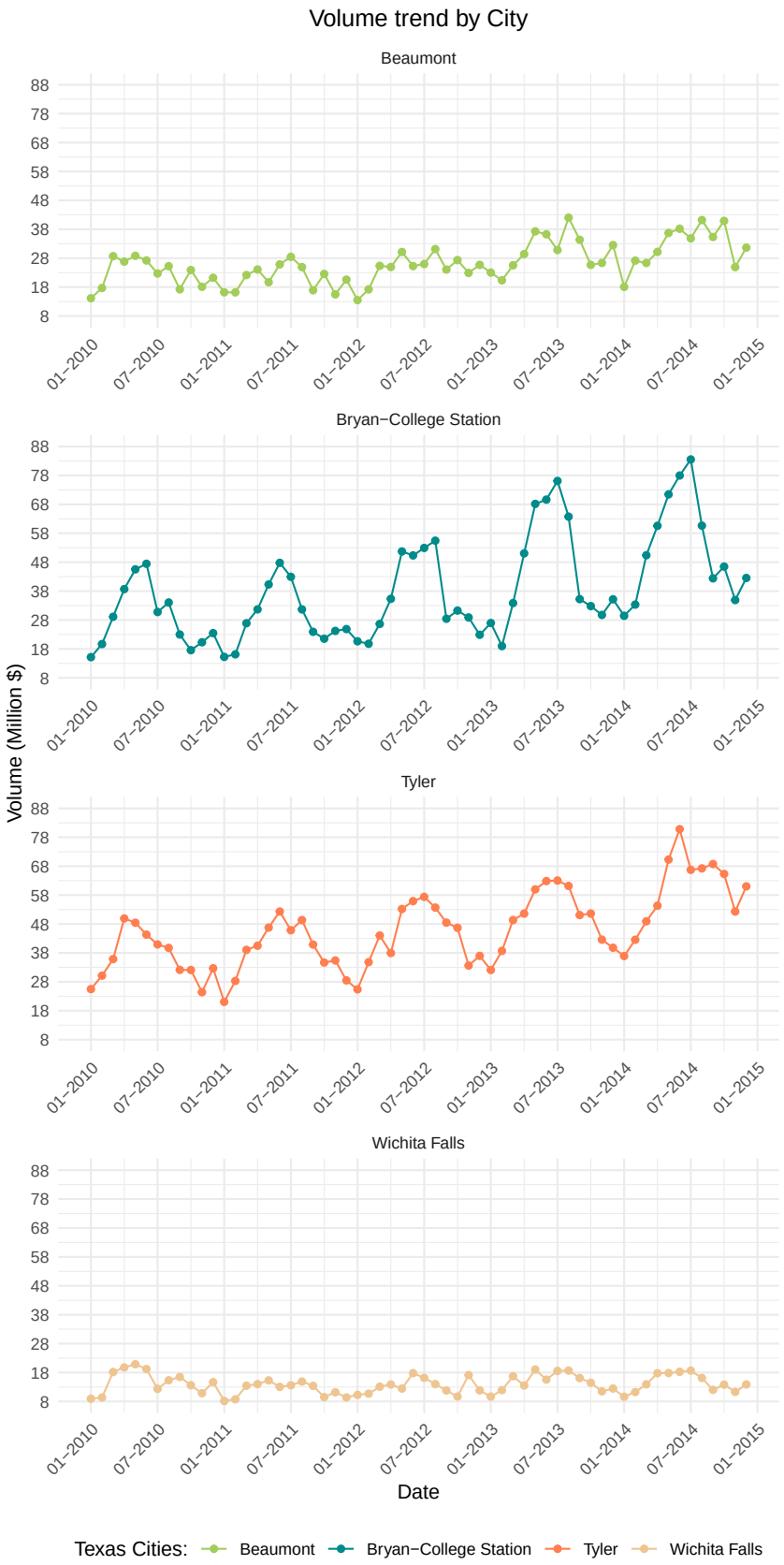


**Figure 15:** Volume comparison by cities over the years.