# Analisi Esplorativa del Mercato Immobiliare del Texas

Riccardo Marotta

2023-07-15

Questo report è il risultato dell'analisi del mercato immobiliare del Texas, svolta utilizzando il software statistico R. E' organizzato in paragrafi corrispondenti alle task di progetto.

## 1. Scarica il dataset realestate_texas.csv da qui e importalo con R, questo contiene dei dati riguardanti le vendite di immobili in Texas

Il dataset realestate_texas.csv è stato importato su R; esso si compone di 240 osservazioni e 8 variabili.

## 2. Indica il tipo di variabili contenute nel dataset

- **city**: variabile qualitativa su scala nominale
- **year**: variabile quantitativa su scala di intervalli
- **month**: variabile qualitativa su scala ordinale, codificata nell'intervallo di valori [1:12]
- **sales**: variabile quantitativa discreta
- **volume**: variabile quantitativa continua
- **median_price**: variabile quantitativa discreta
- **listings**: variabile quantitativa discreta
- **months_inventory**: variabile quantitativa continua

## 3. Calcola Indici di posizione, variabilità e forma per tutte le variabili per le quali ha senso farlo, per le altre crea una distribuzione di frequenza. Commenta tutto brevemente

Il paragrafo è stato organizzato in sezioni corrispondenti alle variabili analizzate.

### 3.1 City

La variabile city è una variabile qualitativa su scala nominale, rappresentativa delle città del Texas presenti nel campione osservato. Considerando che per tale variabile non è possibile calcolare gli indici richiesti, è stata creata la distribuzione di frequenze assolute.

Table 1: **Distribuzione di Frequenze Assolute *(ni)***

| city | ni |
|---|---|
| Beaumont | 60 |
| Bryan-College Station | 60 |
| Tyler | 60 |

| city | ni |
|---|---|
| Wichita Falls | 60 |

Tale distribuzione evidenzia come ciascuna modalità della variabile city sia presente all'interno del campione con la medesima frequenza. Ne consegue che ciascuna modalità corrisponde alla moda della distribuzione (unimodale).

### 3.2 Year

La variabile year è una variabile quantitativa su scala di intervalli, rappresentativa dell'anno di riferimento. Per tale variabile non ha senso il calcolo degli indici richiesti ed è stata pertanto creata la distribuzione di frequenze relative.

Table 2: **Distribuzione di Frequenze Relative *(fi)***

| year | fi |
|---|---|
| 2010 | 0.2 |
| 2011 | 0.2 |
| 2012 | 0.2 |
| 2013 | 0.2 |
| 2014 | 0.2 |

Anche la distribuzione della variabile year è unimodale poiché ciascun valore è presente all'interno del dataset con la medesima frequenza. Inoltre, dalla tabella si può facilmente notare il periodo di riferimento del campione osservato (2010-2014), senza dover ricorrere al calcolo di minimo e massimo.

### 3.3 Month

La variabile month è una variabile qualitativa su scala ordinale, codificata nell'intervallo di valori [1:12], rappresentativa dei mesi dell'anno da Gennaio a Dicembre. Per tale variabile non ha senso il calcolo degli indici richiesti ed è stata pertanto creata la distribuzione di frequenze assolute.

Table 3: **Distribuzione di Frequenze Assolute *(ni)***

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

Anche la distribuzione della variabile month è unimodale dato che ciascuna modalità registra la medesima frequenza nel campione osservato.

Osservando le seguenti distribuzioni di frequenze assolute congiunte delle variabili city, year e month possiamo notare che il dataset si compone di una rilevazione al mese per ogni città e anno di riferimento. Le due tabelle si riferiscono rispettivamente ai mesi di Gennaio e Febbraio, ma il risultato è il medesimo considerando gli altri mesi.

Table 4: **Distribuzione di Frequenze Assolute Congiunte - Gennaio**

| | 2010 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|
| Beaumont | 1 | 1 | 1 | 1 | 1 |
| Bryan-College Station | 1 | 1 | 1 | 1 | 1 |
| Tyler | 1 | 1 | 1 | 1 | 1 |
| Wichita Falls | 1 | 1 | 1 | 1 | 1 |

Table 5: **Distribuzione di Frequenze Assolute Congiunte - Febbraio**

| | 2010 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|
| Beaumont | 1 | 1 | 1 | 1 | 1 |
| Bryan-College Station | 1 | 1 | 1 | 1 | 1 |
| Tyler | 1 | 1 | 1 | 1 | 1 |
| Wichita Falls | 1 | 1 | 1 | 1 | 1 |

### 3.4 Sales

La variabile sales è una variabile quantitativa discreta, rappresentativa del numero totale di vendite. Nella tabella che segue vengono riportati i relativi indici di posizione, variabilità e forma.

Table 6: **Indici della variabile Sales**

| **Min** | **1°Qu.** | **2°Qu.** | **3°Qu.** | **Max** | **Media** | **Moda** | **Range** | **IQR** | **Dev.st** | **CV** | **Asim.** | **Curt.** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 79 | 127 | 175.5 | 247 | 423 | 192.29 | 124 | 344 | 120 | 79.65 | 41.42 | 0.72 | -0.31 |

Relativamente agli **indici di posizione**, il numero totale delle vendite va da un minimo di 79 ad un massimo di 423, con una media di 192.3. Il 1° e 3° quartile sono pari rispettivamente a 127 e 247, mentre il 2° quartile (mediana) è 175.5. Inoltre, il valore di 124 corrisponde alla moda della distribuzione.

Per quanto riguarda gli **indici di variabilità**, vi è una differenza di 344 tra il valore massimo e minimo (range), mentre l'intervallo di variazione del corpo centrale dei dati corrisponde a 120 (IQR). La deviazione standard di 79.6 indica che il numero di vendite spazia mediamente nel seguente intervallo [192.3 - 79.6 ; 192.3 + 79.6]. Il coefficiente di variazione pari a 41.42 indica che la deviazione standard della variabile sales è c.ca il 41% della sua media.

Infine, osservando gli **indici di forma**, l'indice di asimmetria pari a 0.72 evidenzia un'asimmetria positiva, ovvero sono più frequenti valori bassi. L'indice di curtosi pari a -0.31 indica che la distribuzione della variabile sales è platicurtica, ovvero ha una forma più appiattita rispetto a quella della distribuzione normale.

### 3.5 Volume

La variabile volume è una variabile quantitativa continua, rappresentativa del valore totale delle vendite in milioni di dollari. Nella tabella che segue vengono riportati i relativi indici di posizione, variabilità e forma.

Table 7: **Indici della variabile Volume**

| Min | 1°Qu. | 2°Qu. | 3°Qu. | Max | Media | Range | IQR | Dev.st | CV | Asim. | Curt. |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 8.17 | 17.66 | 27.06 | 40.89 | 83.55 | 31.01 | 75.38 | 23.23 | 16.65 | 53.71 | 0.88 | 0.18 |

Relativamente agli **indici di posizione**, i valori di tale variabile vanno da un minimo di 8.17 ad un massimo di 83.55, con una media di c.ca 31. Il 1° e 3° quartile sono pari rispettivamente a 17.66 e 40.89, mentre il 2° quartile (mediana) è 27.06. Poiché i valori della variabile sono su scala continua, è molto difficile che un singolo valore possa ripetersi in modo identico nel campione osservato e, pertanto, non ha molto senso calcolarne la moda.

Per quanto riguarda gli **indici di variabilità**, vi è una differenza di c.ca 75 tra il valore massimo e minimo (range), mentre l'intervallo di variazione del corpo centrale dei dati corrisponde a c.ca 23 (IQR). La deviazione standard di 16.65 indica che i valori della variabile spaziano mediamente nel seguente intervallo [31.01 - 16.65 ; 31.01 + 16.65]. Il coefficiente di variazione pari a 53.71 indica che la deviazione standard della variabile volume è c.ca il 53% della sua media.

Infine, osservando gli **indici di forma**, l'indice di asimmetria pari a 0.88 evidenzia un'asimmetria positiva, ovvero sono più frequenti valori bassi. L'indice di curtosi pari a 0.18 indica che la distribuzione della variabile volume è leptocurtica, ovvero ha una forma più allungata rispetto a quella della distribuzione normale.

### 3.6 Median_price

La variabile median_price è una variabile quantitativa discreta, rappresentativa del prezzo mediano di vendita in dollari. Nella tabella che segue vengono riportati i relativi indici di posizione, variabilità e forma.

Table 8: **Indici della variabile Median_price**

| Min | 1°Qu. | 2°Qu. | 3°Qu. | Max | Media | Moda | Range | IQR | Dev.st | CV | Asim. | Curt. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 73800 | 117300 | 134500 | 150050 | 180000 | 132665 | 130000 | 106200 | 32750 | 22662.1 | 17.08 | -0.36 | -0.62 |

Relativamente agli **indici di posizione**, il prezzo mediano di vendita va da un minimo di 73800 ad un massimo di 180000, con una media di 132665. Il 1° e 3° quartile sono pari rispettivamente a 117300 e 150050, mentre il 2° quartile (mediana) è 134500. Inoltre, il valore di 130000 corrisponde alla moda della distribuzione.

Per quanto riguarda gli **indici di variabilità**, vi è una differenza di 106200 tra il valore massimo e minimo (range), mentre l'intervallo di variazione del corpo centrale dei dati corrisponde a 32750 (IQR). La deviazione standard di 22662.1 indica che il prezzo mediano di vendita spazia mediamente nel seguente intervallo [132665 - 22662.1 ; 132665 + 22662.1]. Il coefficiente di variazione pari a 17.08 indica che la deviazione standard della variabile median_price è c.ca il 17% della sua media.

Infine, osservando gli **indici di forma**, l'indice di asimmetria pari a -0.36 evidenzia un'asimmetria negativa, ovvero sono più frequenti valori alti. L'indice di curtosi pari a -0.62 indica che la distribuzione della variabile median_price è platicurtica, ovvero ha una forma più appiattita rispetto a quella della distribuzione normale.

### 3.7 Listings

La variabile listings è una variabile quantitativa discreta, rappresentativa del numero totale di annunci attivi. Nella tabella che segue vengono riportati i relativi indici di posizione, variabilità e forma.

Table 9: **Indici della variabile Listings**

| Min | 1°Qu. | 2°Qu. | 3°Qu. | Max | Media | Moda | Range | IQR | Dev.st | CV | Asim. | Curt. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 743 | 1026.5 | 1618.5 | 2056 | 3296 | 1738 | 1581 | 2553 | 1029.5 | 752.71 | 43.31 | 0.65 | -0.79 |

Relativamente agli **indici di posizione**, il numero totale di annunci attivi va da un minimo di 743 ad un massimo di 3296, con una media di 1738. Il 1° e 3° quartile sono pari rispettivamente a 1026.5 e 2056, mentre il 2° quartile (mediana) è 1618.5. Inoltre, il valore di 1581 corrisponde alla moda della distribuzione.

Per quanto riguarda gli **indici di variabilità**, vi è una differenza di 2553 tra il valore massimo e minimo (range), mentre l'intervallo di variazione del corpo centrale dei dati corrisponde a 1029.5 (IQR). La deviazione standard di 752.71 indica che il numero totale di annunci attivi spazia mediamente nel seguente intervallo [1738 - 752.71 ; 1738 + 752.71]. Il coefficiente di variazione pari a 43.31 indica che la deviazione standard della variabile listings è c.ca il 43% della sua media.

Infine, osservando gli **indici di forma**, l'indice di asimmetria pari a 0.65 evidenzia un'asimmetria positiva, ovvero sono più frequenti valori bassi. L'indice di curtosi pari a -0.79 indica che la distribuzione della variabile listings è platicurtica, ovvero ha una forma più appiattita rispetto a quella della distribuzione normale.

### 3.8 Months_inventory

La variabile months_inventory è una variabile quantitativa continua e rappresenta la quantità di tempo, espressa in mesi, necessaria per vendere tutte le inserzioni correnti al ritmo attuale delle vendite. Nella tabella che segue vengono riportati i relativi indici di posizione, variabilità e forma.

Table 10: **Indici della variabile Months_inventory**

| Min | 1°Qu. | 2°Qu. | 3°Qu. | Max | Media | Moda | Range | IQR | Dev.st | CV | Asim. | Curt. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3.4 | 7.8 | 8.95 | 10.95 | 14.9 | 9.19 | 8.1 | 11.5 | 3.15 | 2.3 | 25.06 | 0.04 | -0.17 |

Relativamente agli **indici di posizione**, i valori di tale variabile vanno da un minimo di 3.4 ad un massimo di 14.9, con una media di 9.19. Il 1° e 3° quartile sono pari rispettivamente a 7.8 e 10.95, mentre il 2° quartile (mediana) è 8.95. Nonostante i valori della variabile sono su scala continua, non ve ne sono tantissimi e, pertanto, ha senso calcolarne la moda che risulta essere 8.1.

Per quanto riguarda gli **indici di variabilità**, vi è una differenza di c.ca 11.5 tra il valore massimo e minimo (range), mentre l'intervallo di variazione del corpo centrale dei dati corrisponde a c.ca 3.15 (IQR). La deviazione standard di 2.3 indica che i valori spaziano mediamente nel seguente intervallo [9.19 - 2.3 ; 9.19 + 2.3]. Il coefficiente di variazione pari a 25.06 indica che la deviazione standard della variabile months_inventory è c.ca il 25% della sua media.

Infine, osservando gli **indici di forma**, l'indice di asimmetria di 0.04 evidenzia una quasi perfetta simmetria della distribuzione. L'indice di curtosi pari a -0.17 indica che la distribuzione della variabile months_inventory è platicurtica, ovvero ha una forma più appiattita rispetto a quella della distribuzione normale.

## 4. Qual è la variabile con variabilità più elevata? Come ci sei arrivato? E quale quella più asimmetrica?

La variabile **volume** è quella con **variabilità più elevata** in quanto, tra tutte le variabili, ha il coefficiente di variazione più grande (53.71). Inoltre, tale variabile risulta anche quella **più asimmetrica** poiché presenta un indice di asimmetria più grande in valore assoluto (0.88).

## 5. Dividi una delle variabili quantitative in classi, scegli tu quale e come, costruisci la distribuzione di frequenze, il grafico a barre corrispondente e infine calcola l'indice di Gini.

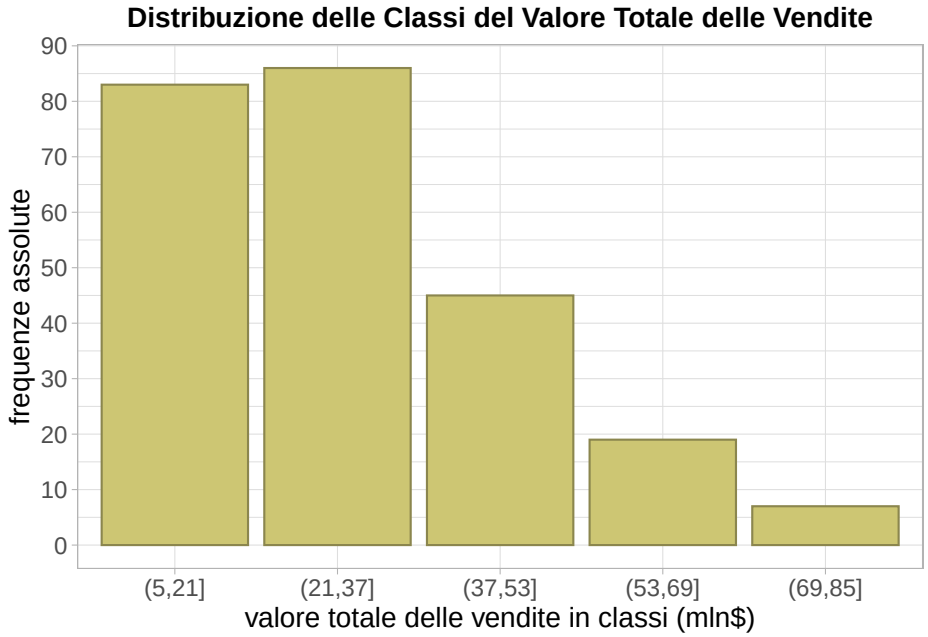
E' stata presa come riferimento la variabile volume, dividendola in 5 classi di uguale ampiezza e sono state costruite le distribuzioni di frequenze (assolute: *ni*, relative: *fi*, cumulate: *Ni*, relative cumulate: *Fi*).

Table 11: **Distribuzioni di Frequenze delle Classi del Valore Totale delle Vendite (mln$)**

| **volume** | ni | fi | Ni | Fi |
|---|---|---|---|---|
| (5,21] | 83 | 0.35 | 83 | 0.35 |
| (21,37] | 86 | 0.36 | 169 | 0.70 |
| (37,53] | 45 | 0.19 | 214 | 0.89 |
| (53,69] | 19 | 0.08 | 233 | 0.97 |
| (69,85] | 7 | 0.03 | 240 | 1.00 |

Dalla distribuzione di frequenze assolute *(ni)* si può notare come la maggior parte delle osservazioni ricada nelle classi più basse (5,21] e (21,37], confermando il risultato emerso dall'indice di asimmetria dei dati originari (0.88). Infatti, dalla distribuzione di frequenze relative cumulate *(Fi)* si osserva come tali classi rappresentino il 70% del campione osservato. La classe centrale (37,53] registra una frequenza relativa *(fi)* del 19%. Ne consegue che solamente l'11% delle rilevazioni osservate nel campione presenta un valore totale delle vendite superiore ai 53 milioni di dollari.

Si riporta di seguito il grafico a barre della distribuzione di frequenze assolute, dal quale emerge facilmente come le classi più basse della variabile volume siano le più frequenti.




Inoltre, l'**indice di Gini** della variabile volume suddivisa in classi è pari a 0.88. Tale valore indica che la distribuzione è molto eterogenea e che le unità statistiche non si concentrano soltanto in una classe.

## 6. Indovina l'indice di gini per la variabile city

Poiché ciascuna modalità della variabile city registra la medesima frequenza, si può dedurre che il relativo indice di Gini è uguale a 1. Tale variabile risulta pertanto caratterizzata da un'eterogeneità massima.

## 7. Qual è la probabilità che presa una riga a caso di questo dataset essa riporti la città "Beaumont"? E la probabilità che riporti il mese di Luglio? E la probabilità che riporti il mese di dicembre 2012?

La probabilità che, presa una riga a caso di questo dataset, essa riporti la città "Beaumont" è pari al 25% in quanto vi sono 4 modalità per la variabile city, tutte con la medesima frequenza (n°casi favorevoli = 1 / n°casi possibili = 4).

La probabilità che, presa una riga a caso di questo dataset, essa riporti il mese di Luglio è pari all'8% c.ca in quanto tutti i mesi dell'anno presentano la medesima frequenza (n°casi favorevoli = 1 / n°casi possibili = 12).

La probabilità che, presa una riga a caso di questo dataset, essa riporti il mese di dicembre 2012 è c.ca l'1% in quanto dalla distribuzione di frequenze assolute congiunte delle variabili month e year emergono 4 rilevazioni, una per ciascuna città (n°casi favorevoli = 4 / n°casi possibili = 240).

## 8. Esiste una colonna col prezzo mediano, creane una che indica invece il prezzo medio, utilizzando le altre variabili che hai a disposizione.

A partire dalle variabili volume e sales, è stata creata una nuova colonna nel dataset denominata mean_price e rappresentativa del prezzo medio di vendita in dollari. E' stata calcolata come rapporto tra la variabile volume (trasformata da milioni di dollari a dollari) e sales.

## 9. Prova a creare un'altra colonna che dia un'idea di "efficacia" degli annunci di vendita. Riesci a fare qualche considerazione?

Fatte le seguenti assunzioni:

- ogni immobile sul mercato deve avere il relativo annuncio
- la variabile listings, rappresentativa del numero totale di annunci attivi, non tiene conto degli annunci relativi agli immobili venduti
- gli annunci scaduti riguardano solamente quelli relativi ad immobili venduti

è stata creata una nuova variabile denominata sales_rate. Tale colonna è stata calcolata come rapporto tra la variabile sales e il numero totale di annunci (attivi e scaduti) e rappresenta il tasso di successo degli annunci di vendita, fornendone pertanto un'indicazione di "efficacia".

Nel campione osservato si registra un tasso di successo degli annunci di vendita in media del 10%, con una deviazione standard pari al 3% c.ca.

Inoltre, è stato calcolato il coefficiente di correlazione di Pearson tra la nuova variabile sales_rate e la variabile months_inventory, rappresentativa della quantità di tempo, espressa in mesi, necessaria per vendere tutte le inserzioni correnti al ritmo attuale delle vendite. La correlazione tra le due variabili è di -0.67 evidenziando una relazione lineare negativa piuttosto forte. Ne consegue che all'aumentare del tasso di successo degli annunci di vendita, la quantità in mesi necessaria per vendere tutte le inserzioni tende a diminuire (e viceversa).

## 10. Prova a creare dei summary(), o semplicemente media e deviazione standard, di alcune variabili a tua scelta, condizionatamente alla città, agli anni e ai mesi.

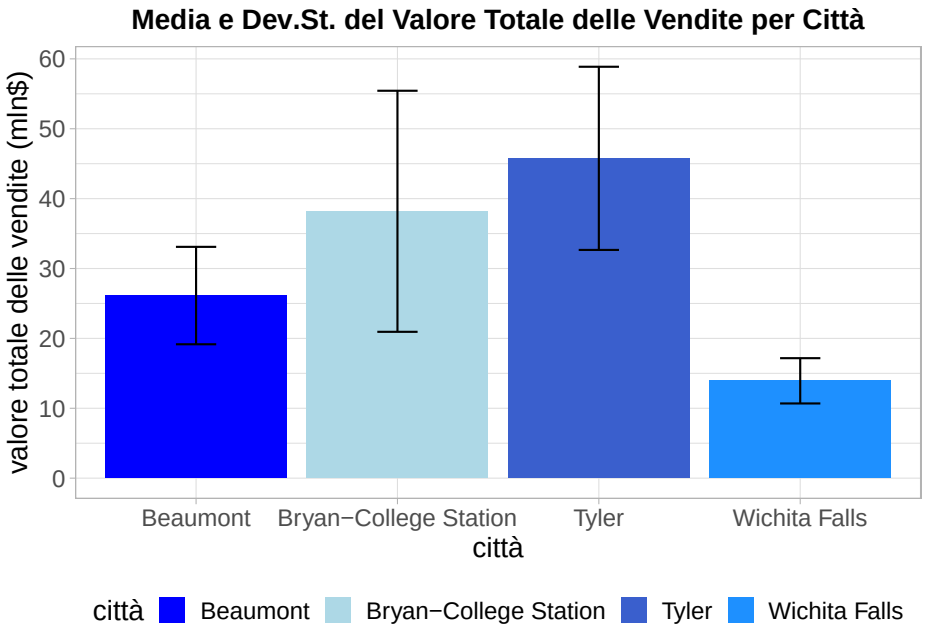
Il paragrafo è stato organizzato in sezioni corrispondenti alle variabili analizzate.

### 10.1 Volume

Prendendo come riferimento la variabile volume, la tabella seguente e il relativo grafico a barre ne rappresentano media e deviazione standard condizionatamente alla variabile city. Nel grafico l'altezza di ciascuna barra identifica la media della variabile volume condizionata ad una determinata città, mentre la linea verticale posta su ciascuna barra rappresenta la relativa deviazione standard. Si può notare come la città di Tyler sia quella con una media del valore totale delle vendite più alta, di oltre il triplo rispetto a quanto registrato nella città di Wichita Falls. Inoltre, Bryan-College Station è la città con la deviazione standard più alta, mentre Wichita Falls e Beaumont presentano i valori più bassi.

Table 12: **Media e Deviazione Standard del Valore Totale delle Vendite per Città**

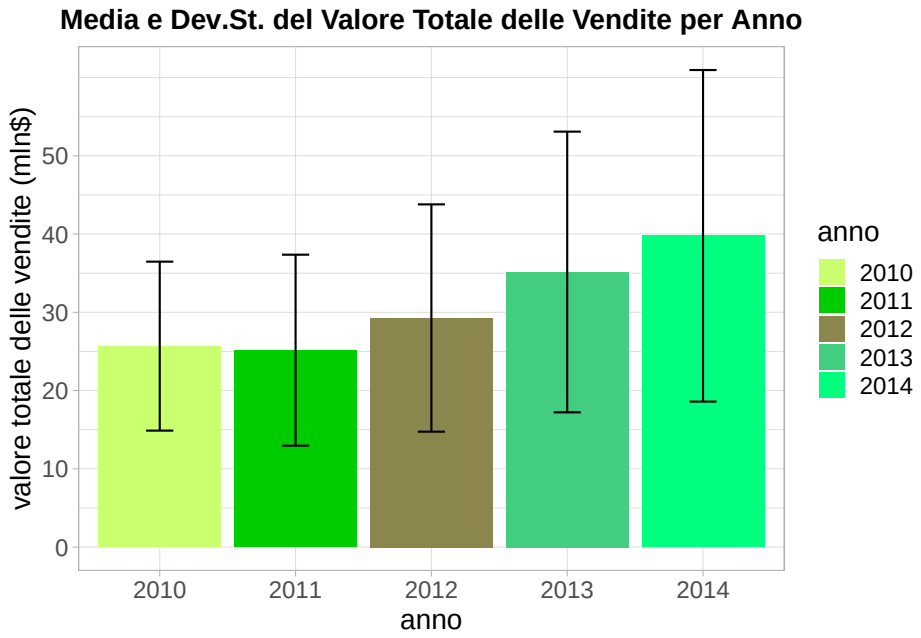
| city | media | dev.st |
|---|---|---|
| Beaumont | 26.13 | 6.97 |
| Bryan-College Station | 38.19 | 17.25 |
| Tyler | 45.77 | 13.11 |
| Wichita Falls | 13.93 | 3.24 |




La tabella seguente e il relativo grafico a barre rappresentano media e deviazione standard della variabile volume condizionatamente alla variabile year. Si può facilmente notare come, al crescere degli anni, corrisponda un aumento del valore totale delle vendite sia in media che in termini di deviazione standard. Infatti, il 2014 ha fatto registrare in media un valore totale delle vendite di quasi 40 mln, di oltre 10 mln rispetto al primo anno del periodo di riferimento.

Table 13: **Media e Deviazione Standard del Valore Totale delle Vendite per Anno**

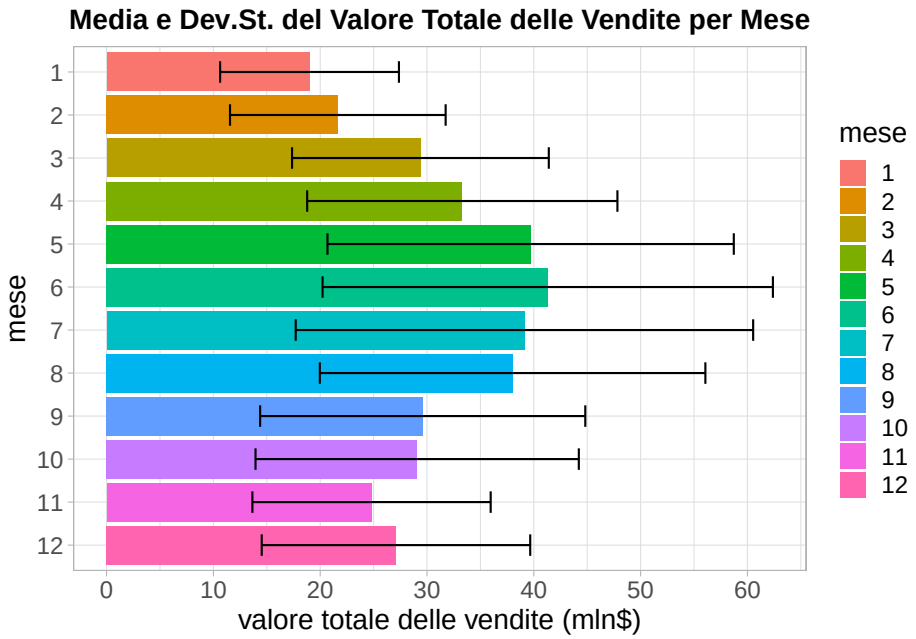
| year | media | dev.st |
|---|---|---|
| 2010 | 25.68 | 10.80 |
| 2011 | 25.16 | 12.20 |
| 2012 | 29.27 | 14.52 |
| 2013 | 35.15 | 17.93 |
| 2014 | 39.77 | 21.19 |



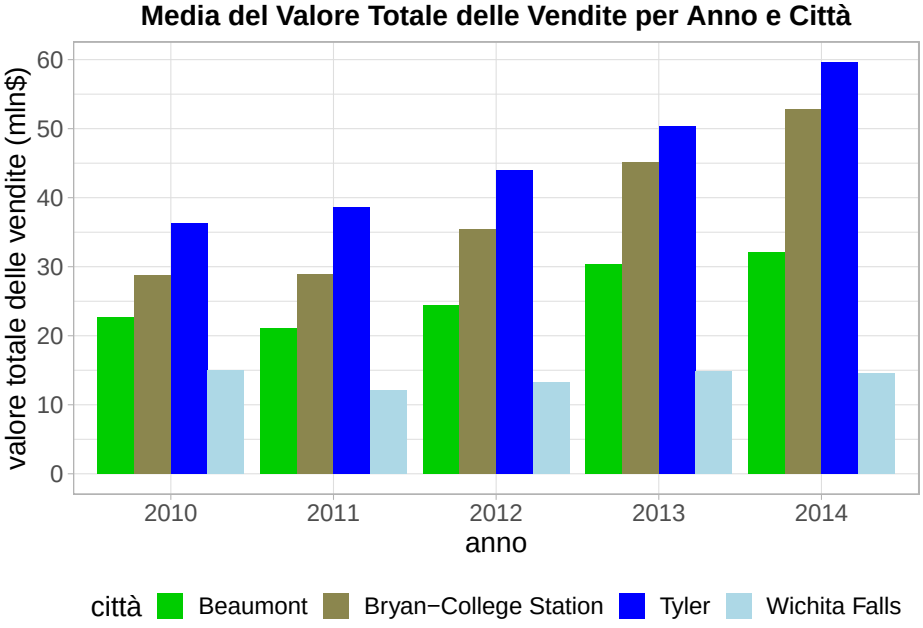

La tabella seguente e il relativo grafico a barre orizzontali rappresentano media e deviazione standard della variabile volume condizionatamente alla variabile month. A differenza dei precedenti grafici, in questo caso la lunghezza di ciascuna barra identifica la media della variabile volume condizionata ad un determinato mese, mentre la linea orizzontale posta su ciascuna barra rappresenta la relativa deviazione standard. Si può notare come i mesi di Maggio, Giugno, Luglio e Agosto siano quelli contraddistinti da valori totali delle vendite più alti sia in media (>38 mln) che in termini di deviazione standard (>18 mln). Viceversa, Gennaio è il mese che registra i volumi delle vendite più bassi con una media di 19 mln e una deviazione standard di 8 mln.

Table 14: **Media e Deviazione Standard del Valore Totale delle Vendite per Mese**

| month | media | dev.st |
|---|---|---|
| 1 | 19.00 | 8.37 |
| 2 | 21.65 | 10.09 |
| 3 | 29.38 | 12.02 |
| 4 | 33.30 | 14.52 |
| 5 | 39.70 | 19.02 |
| 6 | 41.30 | 21.08 |
| 7 | 39.12 | 21.41 |
| 8 | 38.01 | 18.05 |
| 9 | 29.60 | 15.22 |
| 10 | 29.08 | 15.13 |
| 11 | 24.81 | 11.15 |
| 12 | 27.09 | 12.57 |




Inoltre, per avere una maggior granularità della variabile volume, è stata condizionata congiuntamente alle variabili year e city. Il grafico a barre verticali seguente rappresenta la media del valore totale delle vendite per anno e città ed è utile per effettuare confronti sia tra le diverse città in un determinato anno, sia considerando una medesima città in diversi anni. Si può notare come la città di Tyler abbia registrato in ogni anno volumi in media sempre superiori ai 35 mln, fino a sfiorare la soglia dei 60 mln nell'ultimo anno di riferimento. D'altro canto, si registra un andamento abbastanza stazionario della media del valore totale delle vendite nella città di Wichita Falls, con valori compresi tra c.ca 12 mln e 15 mln nei diversi anni. Inoltre, in ogni anno si può osservare come la "classifica" della media dei volumi delle vendite sia stata sempre la medesima, con la città di Tyler con i valori più alti, seguita in ordine via via decrescente dalle città di Bryan-College Station, Beaumont e Wichita Falls.

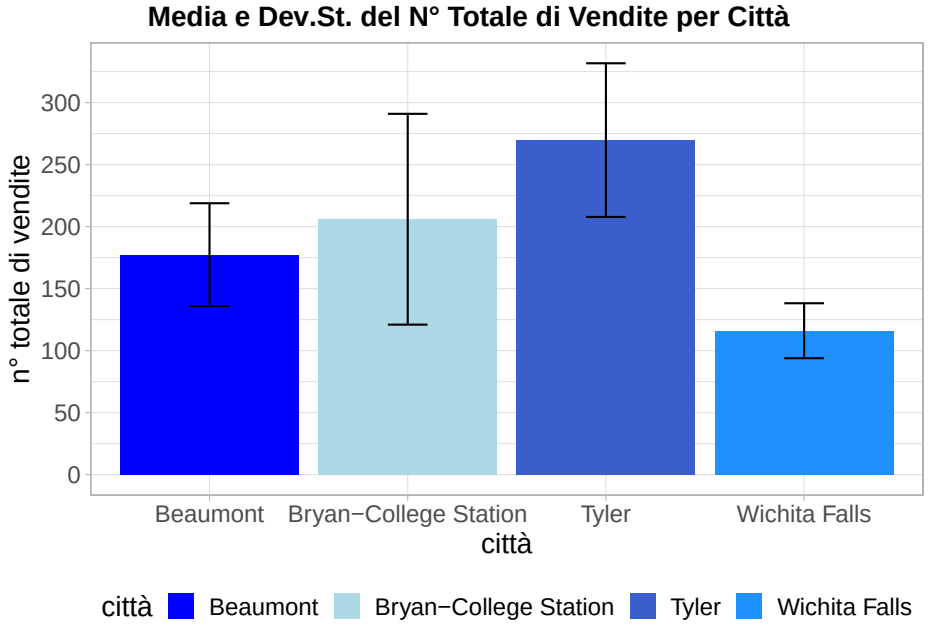


### 10.2 Sales

Prendendo come riferimento la variabile sales, la tabella seguente e il relativo grafico a barre ne rappresentano media e deviazione standard condizionatamente alla variabile city. Si può notare come la città di Tyler sia quella con una media del numero totale delle vendite più alta, di oltre il doppio rispetto a quanto registrato nella città di Wichita Falls. Inoltre, Bryan-College Station è la città che registra la deviazione standard maggiore, mentre Wichita Falls quella minore.

Table 15: **Media e Deviazione Standard del N° Totale di Vendite per Città**

| city | media | dev.st |
|---|---|---|
| Beaumont | 177.38 | 41.48 |
| Bryan-College Station | 205.97 | 84.98 |
| Tyler | 269.75 | 61.96 |
| Wichita Falls | 116.07 | 22.15 |

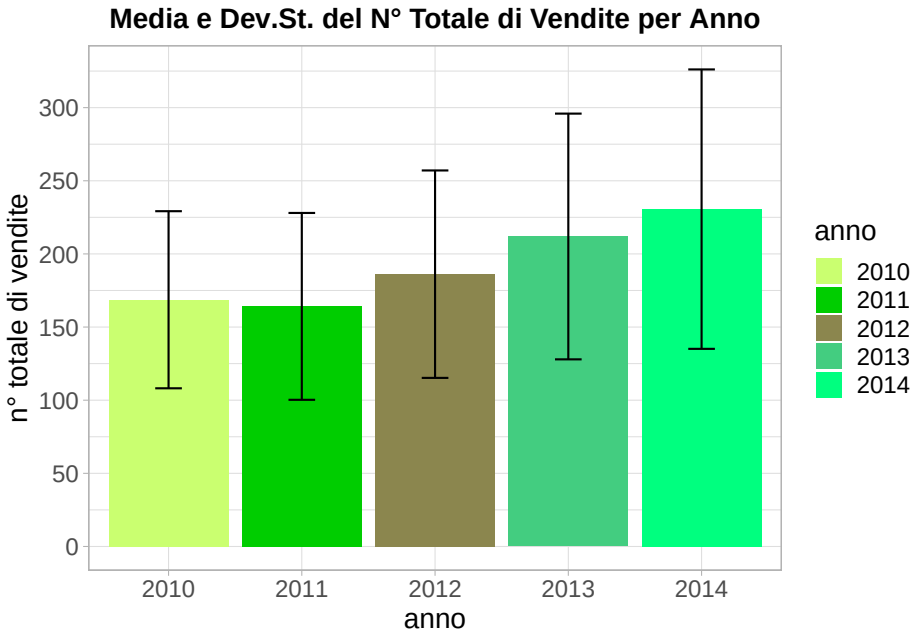



La tabella seguente e il relativo grafico a barre rappresentano media e deviazione standard della variabile sales condizionatamente alla variabile year. Si può facilmente notare come, al crescere degli anni, corrisponda quasi sempre un aumento del numero totale delle vendite sia in media che in termini di deviazione standard. Infatti, negli ultimi due anni (2013-2014) si sono registrati in media un numero totale di vendite maggiore (oltre le 210 unità), mentre nei primi due anni (2010-2011) i valori più bassi (sulle 160 unità).

Table 16: **Media e Deviazione Standard del N° Totale di Vendite per Anno**

| year | media | dev.st |
|---|---|---|
| 2010 | 168.67 | 60.54 |
| 2011 | 164.12 | 63.87 |
| 2012 | 186.15 | 70.91 |

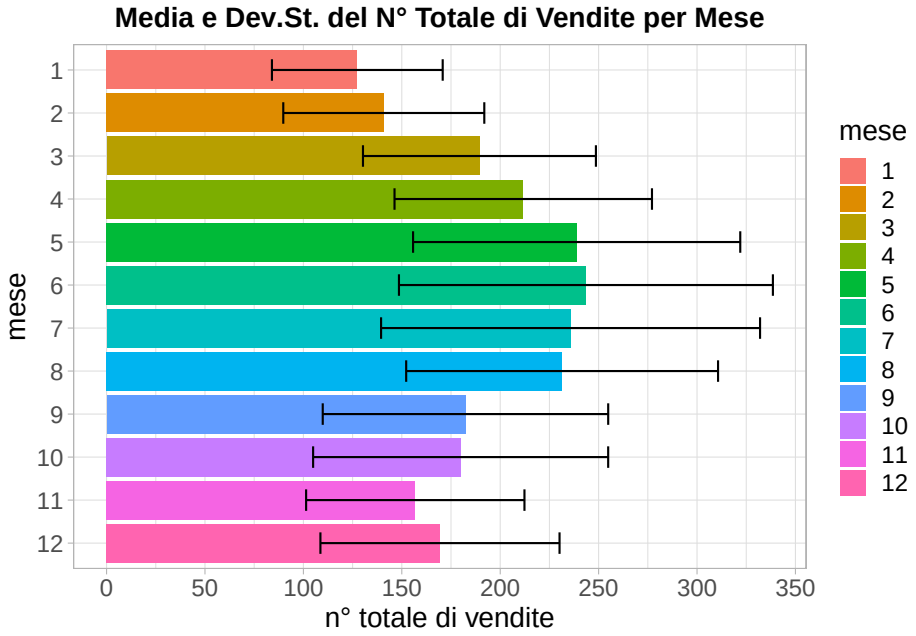
| year | media | dev.st |
|---|---|---|
| 2013 | 211.92 | 84.00 |
| 2014 | 230.60 | 95.51 |



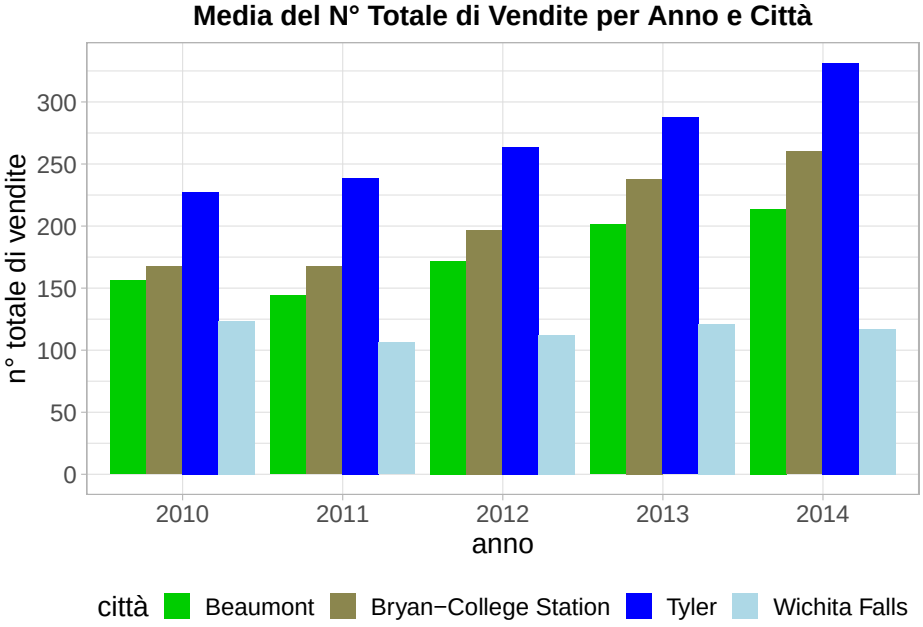

La tabella seguente e il relativo grafico a barre orizzontali rappresentano media e deviazione standard della variabile sales condizionatamente alla variabile month. Come per la media e deviazione standard della variabile volume condizionata a month, anche in questo caso i mesi di Maggio, Giugno, Luglio e Agosto sono quelli contraddistinti da valori maggiori sia in media (>230 unità) che in termini di deviazione standard (>79 unità), mentre Gennaio risulta ancora una volta il mese "peggiore".

Table 17: **Media e Deviazione Standard del N° Totale di Vendite per Mese**

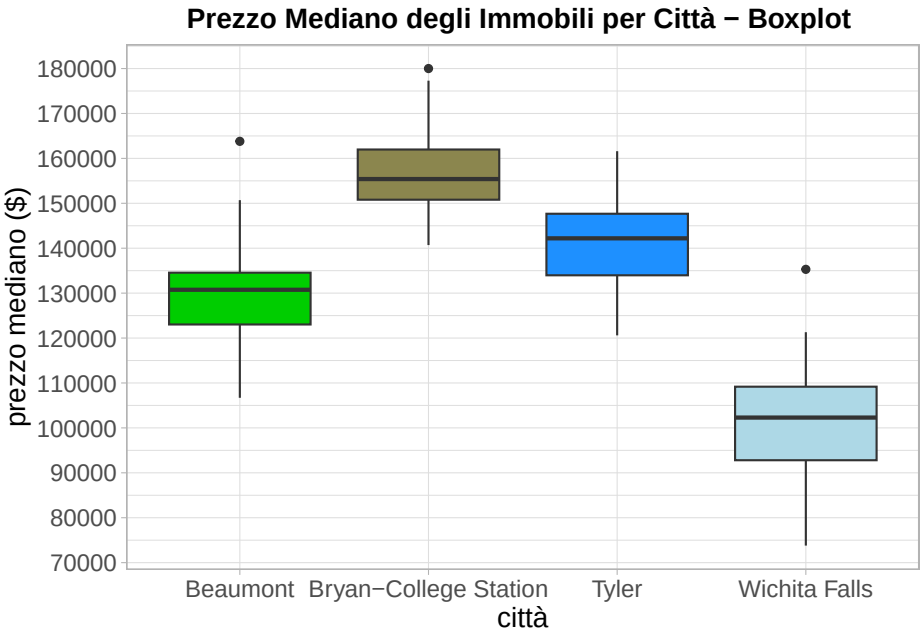
| month | media | dev.st |
|---|---|---|
| 1 | 127.40 | 43.38 |
| 2 | 140.85 | 51.07 |
| 3 | 189.45 | 59.18 |
| 4 | 211.70 | 65.40 |
| 5 | 238.85 | 83.12 |
| 6 | 243.55 | 95.00 |
| 7 | 235.75 | 96.27 |
| 8 | 231.45 | 79.23 |
| 9 | 182.35 | 72.52 |
| 10 | 179.90 | 74.95 |
| 11 | 156.85 | 55.47 |
| 12 | 169.40 | 60.75 |




Anche la variabile sales è stata condizionata congiuntamente alle variabili year e city. Il grafico seguente consente di confrontare la media del numero totale delle vendite sia tra le diverse città in un determinato anno, sia in una medesima città su diversi anni. Si può notare come, in ogni anno, la città di Tyler abbia registrato in media un numero totale di vendite sempre superiore alle 200 unità, con oltre 300 unità di immobili venduti nell'ultimo anno di riferimento. D'altro canto, Wichita Falls presenta un andamento abbastanza stazionario della media del numero totale delle vendite, con valori compresi tra le 100 e le 125 unità. La città di Beaumont ha fatto registrare una leggera diminuzione della media delle unità vendute nell'anno successivo al 2010, per poi osservare degli aumenti negli anni successivi, raggiungendo dal 2013 la soglia delle 200 unità. Infine, nella città di Bryan-College Station si rilevano valori della media delle unità vendute molto alti negli ultimi due anni (>225 unità), rispetto alle c.ca 160 unità degli anni 2010-2011.



## 11. Utilizza i boxplot per confrontare la distribuzione del prezzo mediano delle case tra le varie città. Commenta il risultato




Dal grafico emerge subito che la città di Bryan-College Station presenta i valori più alti per il prezzo mediano degli immobili (range c.ca 140k-180k), mente la città di Wichita Falls i valori più bassi (range c.ca 75k-135k).

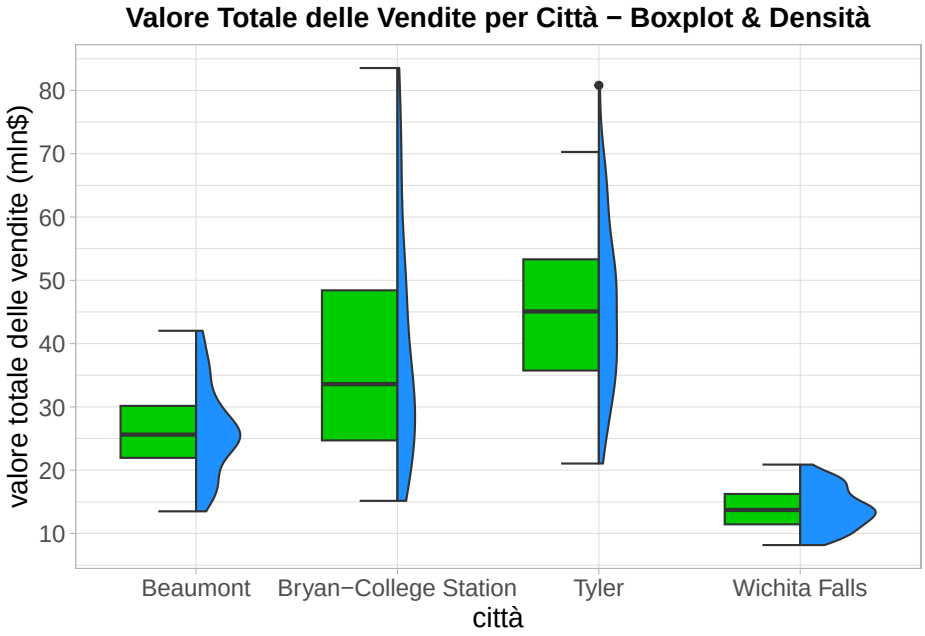
Considerando la mediana del prezzo mediano per ciascuna città, individuata dalla linea centrale nella scatola, si registra il valore più alto nella città di Bryan-College Station (c.ca 155k). Le restanti città di Tyler, Beaumont e Wichita Falls presentano rispettivamente un valore di c.ca 140k, 130k e 100k.

Considerando la scatola di ciascun boxplot, ovvero la differenza interquartile, si può notare come questa sia più ampia nella città di Wichita Falls. Ne consegue che per tale città vi è una maggior variabilità del prezzo mediano all'interno del 50% dei valori compresi tra il 1° e 3° quartile.

Inoltre, si può notare come la città di Tyler sia l'unica a non registrare alcun outlier, mentre le città di Beaumont e Wichita Falls presentano degli outliers molto distanti dai rispettivi massimi relativi, individuati dalla parte finale delle linee uscenti verso l'alto dalle scatole. Invece, per la città di Bryan-College Station l'outlier non è molto distante dal suo massimo relativo.

## 12. Utilizza i boxplot o qualche variante per confrontare la distribuzione del valore totale delle vendite tra le varie città ma anche tra i vari anni. Qualche considerazione da fare?

In entrambi i grafici presenti nel paragrafo è possibile visualizzare per ciascuna modalità/valore della variabile city/year, mezzo boxplot a sinistra e mezza densità di probabilità a destra per la variabile volume.

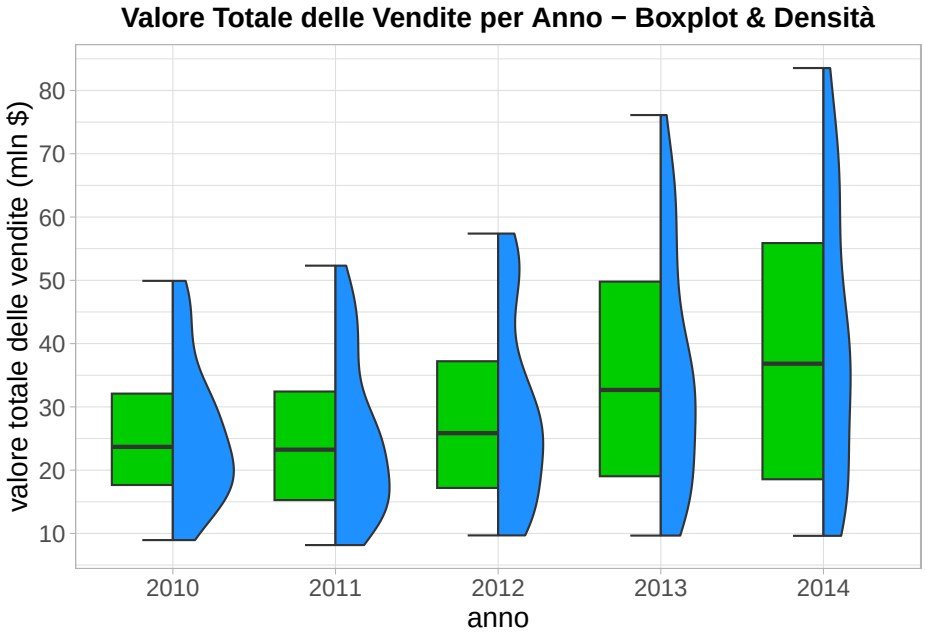



Considerando i **boxplot** sul lato sinistro di ciascuna modalità della variabile city, è possibile notare come la città di Tyler presenti il valore totale delle vendite maggiore rispetto alle altre città in termini di mediana (c.ca 45 mln), mentre la città di Wichita Falls il valore più basso (c.ca 13 mln). Le città di Bryan-College Station e Beaumont registrano rispettivamente una mediana di c.ca 33 mln e 25 mln. Considerando la scatola di ciascun boxplot, ovvero la differenza interquartile, si può notare come questa sia più ampia nella città di Bryan-College Station, con una conseguente variabilità maggiore all'interno del 50% dei valori compresi tra il 1° e 3° quartile (c.ca 25-48 mln). Inoltre, tale città registra il range di valori più ampio, da un minimo di 15 mln ad un massimo di c.ca 83 mln. La città di Tyler presenta un outlier di c.ca 80 mln distante di c.ca 10 mln dal suo massimo relativo. Infine, la città di Wichita Falls risulta essere quella con una differenza interquartile e con un range minore.

Considerando le **densità di probabilità** sul lato destro di ciascuna modalità, si osserva la forma delle distribuzioni di probabilità della variabile volume condizionata ad una determinata città. Tutte le città manifestano dei valori bassi più frequenti e, pertanto, possiamo concludere che le relative distribuzioni sono asimmetriche positive. Ciò è particolarmente evidente nella città di Bryan-College Station. Inoltre, le distribuzioni di probabilità della variabile volume condizionate alle città di Beaumont, Tyler e Wichita Falls presentano una forma più appiattita rispetto alla normale e possiamo concludere che esse sono platicurtiche. Le prime evidenze grafiche sulla forma delle distribuzioni di probabilità condizionate sono confermate dai relativi indici mostrati nella seguente tabella.

Table 18: **Valore Totale delle Vendite per Città - Indici di forma**

| city | asimmetria | curtosi |
|---|---|---|
| Beaumont | 0.36 | -0.36 |
| Bryan-College Station | 0.86 | -0.08 |
| Tyler | 0.35 | -0.39 |
| Wichita Falls | 0.19 | -0.90 |



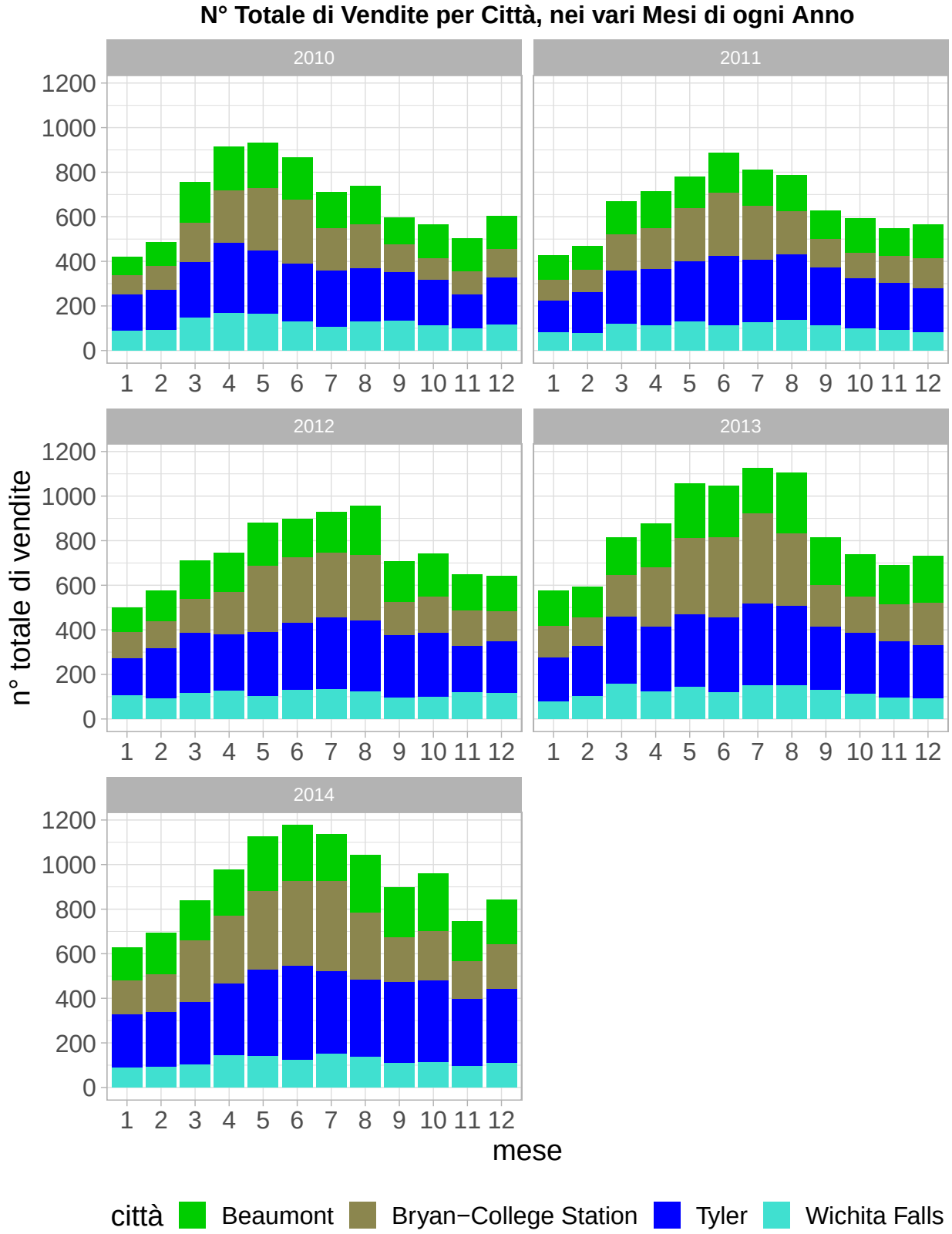

Considerando i **boxplot** sul lato sinistro di ogni anno, si può osservare come il minimo del valore totale delle vendite sia abbastanza stazionario, con valori intorno ai 10 mln. Viceversa, il massimo della variabile volume condizionata manifesta un trend positivo, passando da 50 mln nel 2010 a quasi 85 mln nell'ultimo anno di riferimento. La mediana registra un leggero aumento negli ultimi due anni. Infatti, nei primi tre anni essa oscilla intorno ai 25 mln, mentre nel 2013 e 2014 essa si attesta rispettivamente su c.ca 33 mln e 37 mln. Inoltre, negli ultimi due anni si può notare come la scatola rappresentativa della differenza interquartile sia più ampia rispetto ai primi tre anni. Ad esempio, nel 2014 essa è compresa nell'intervallo di c.ca 18-56 mln. Infine, non sono presenti outliers nei diversi anni e quindi il massimo coincide sempre con il massimo relativo.

Considerando le **densità di probabilità** sul lato destro di ogni anno, si può osservare la forma delle distribuzioni di probabilità della variabile volume condizionata ai diversi anni. Tutte le distribuzioni presentano valori bassi più frequenti e hanno una forma più appiattita rispetto alla normale. Pertanto, possiamo concludere che esse hanno un'asimmetria positiva e che sono platicurtiche. Le prime evidenze grafiche sulla forma di tali distribuzioni sono confermate dai relativi indici mostrati nella seguente tabella.

Table 19: **Valore Totale delle Vendite per Anno - Indici di forma**

| year | asimmetria | curtosi |
|---|---|---|
| 2010 | 0.62 | -0.44 |
| 2011 | 0.61 | -0.69 |
| 2012 | 0.54 | -0.88 |
| 2013 | 0.54 | -0.71 |
| 2014 | 0.37 | -0.95 |

**13. Usa un grafico a barre sovrapposte per ogni anno, per confrontare il totale delle vendite nei vari mesi, sempre considerando le città. Prova a commentare ciò che viene fuori. Già che ci sei prova anche il grafico a barre normalizzato.**



Il **grafico a barre** con il numero totale di vendite per ciascuna città nei vari mesi, è stato costruito nella pagina precedente spacchettandolo in cinque diversi grafici per ogni anno di riferimento (2010-2014).

Si può notare come, al netto del 2010, le vendite maggiori di immobili tendano a concentrarsi in ciascun anno nei mesi da Maggio ad Agosto. Durante gli ultimi due anni del periodo di riferimento, si registra un aumento complessivo nelle vendite, con i mesi di Maggio, Giugno, Luglio e Agosto che registrano in entrambi gli anni delle vendite superiori alle 1000 unità. In ogni anno Gennaio risulta il mese con il numero più basso di immobili venduti.

Inoltre, in ogni anno si può notare come nei mesi contraddistinti da numeri più alti nelle vendite, le città di Tyler e Bryan-College Station tendano ad equivalersi, mentre nei mesi "peggiori" Tyler registra un numero di vendite superiore rispetto a Bryan-College Station.

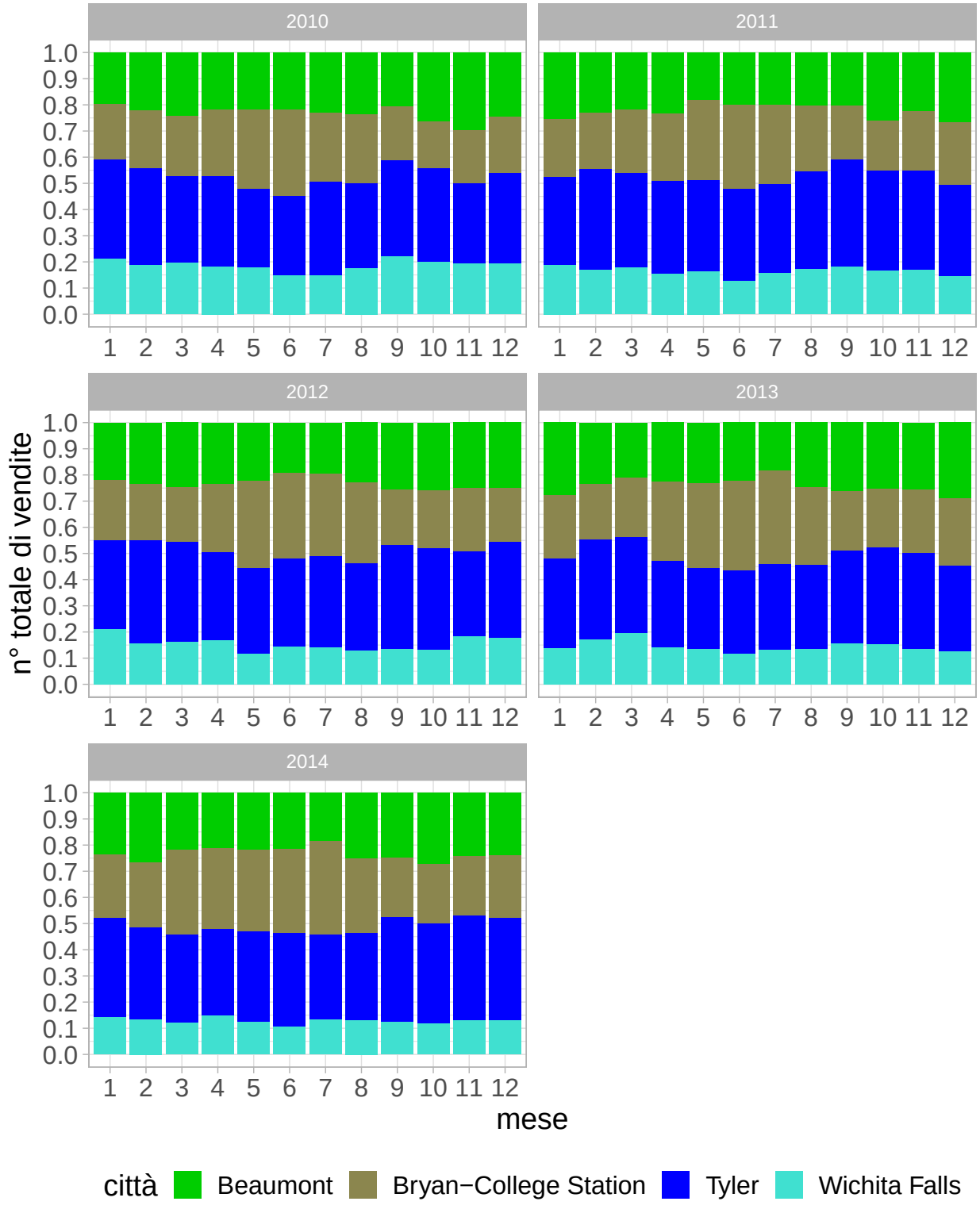
Infine, considerando ciascun mese di ogni anno, risulta evidente come la città di Wichita Falls non sia mai riuscita a raggiungere le 200 unità di immobili venduti.

A pagina seguente viene riportato anche il relativo **grafico normalizzato** dal quale è possibile fare ulteriori confronti, grazie alle frequenze relative rappresentate sull'asse delle y.

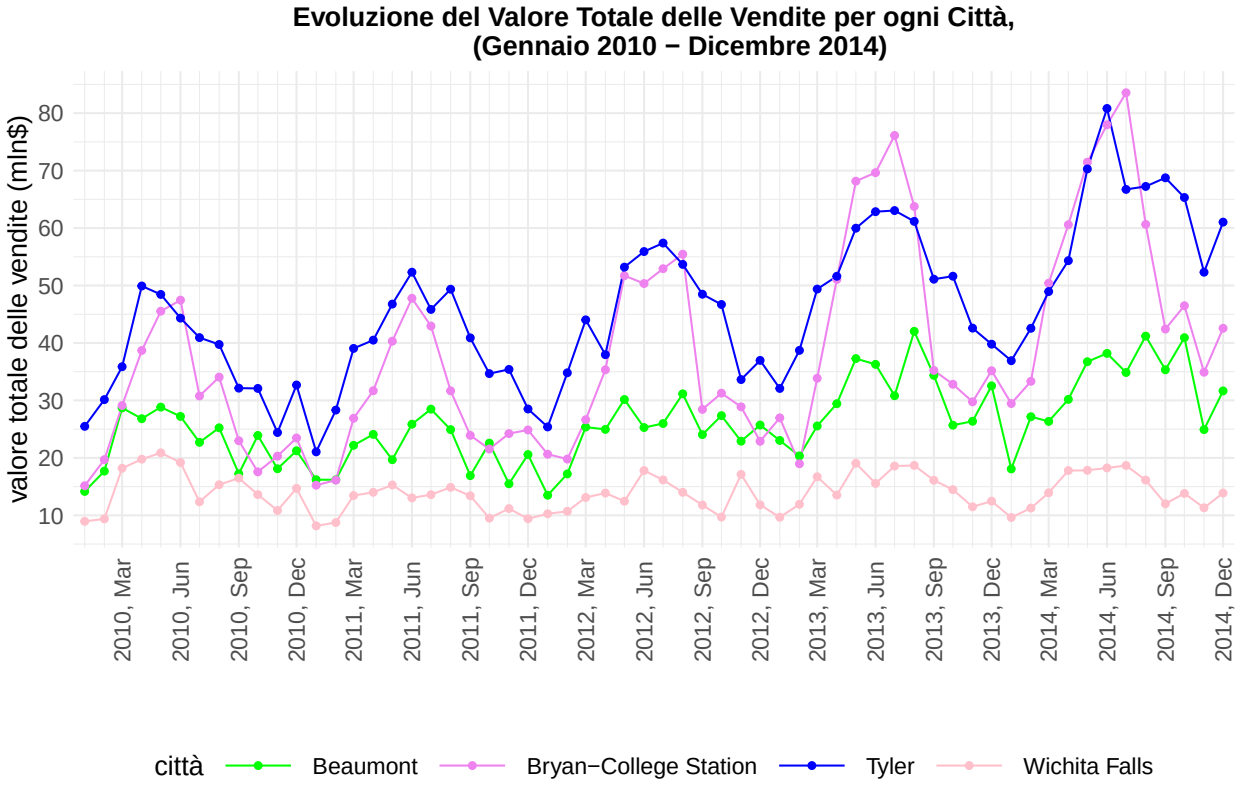
Complessivamente, considerando ciascun mese di ogni anno, si può notare come la città di Tyler registri sempre un numero totale di vendite mai inferiore al 30% delle osservazioni. Inoltre, considerando ad esempio il 2014, la frequenza relativa delle vendite attribuite a tale città si attesta su c.ca il 40% nei mesi di Gennaio, Settembre, Novembre e Dicembre.

Infine, in quasi tutti i mesi di ogni anno, le città di Beaumont e Wichita Falls rappresentano rispettivamente il 20-25% e 10-15% delle vendite osservate nel campione.

**N° Totale di Vendite per Città, nei vari Mesi di ogni Anno − Grafico Normalizzato**



**14. Crea un line chart di una variabile a tua scelta per fare confronti commentati fra città e periodi storici. Consigli: Prova inserendo una variabile per volta. Prova a usare variabili esterne al dataset, tipo vettori creati da te appositamente.**




Il grafico mostra l'evoluzione della variabile volume, per ciascuna città, nel periodo di riferimento osservato nel campione (Gennaio 2010 - Dicembre 2014). Ogni punto della linea spezzata individua una rilevazione mensile per un totale di 60 osservazioni per ogni città.

Considerando l'intero periodo, si può notare come la città di Bryan-College Station abbia fatto registrare, a Luglio 2014, il valore totale più alto totale delle vendite (quasi 85 mln). Tuttavia, in ben 53 mesi su 60, la città di Tyler presenta i valori nelle vendite più alti rispetto a tutte le altre città, Bryan-College Station compresa. D'altro canto, il valore più basso nelle vendite si è registrato a Gennaio 2011 nella città di Wichita Falls, la quale presenta i valori più bassi rispetto alle altre città in tutte le rilevazioni mensili.

In relazione agli ultimi tre anni del periodo di riferimento, il mese di Luglio risulta sempre quello contraddistinto da un valore totale delle vendite maggiore.

Inoltre, in ogni anno, si osserva nei mesi centrali una notevole differenza tra le città di Tyler e Bryan-College Station, caratterizzate da volumi maggiori, e le città di Beaumont e Wichita Falls che presentano volumi inferiori.