Analisi del mercato immobiliare del Texas

Il dataset realestate_texas.csv è costituito da 240 righe di 8 variabili:

- **City:**

variabile qualitativa su scala nominale, non è possibile stabilire un ordine di grandezza (minore o maggiore) ma solamente un ordine di uguaglianza (a=b o a≠b)

- **Year:**

Variabile qualitativa ordinale. Nonostante si riferisca al tempo, nel dataset di riferimento assume solamente 5 modalità (2010, 11, 12, 13 e 14) ordinabili secondo ordine di grandezza

- **Month:**

Variabile qualitativa ordinale codificata. Si riferisce al tempo in mesi che può assumere 12 modalità. Nel dataset, il valore del mese viene assegnato un valore numerico corrispondente al mese di riferimento (1 = gennaio, 2 = febbraio, 3 = marzo, etc.),

inoltre, la variabile è ordinabile con un ordine di grandezza (gennaio viene prima di febbraio che viene prima di marzo e così via).

- **Sales:**

Variabile quantitativa discreta. Il numero delle vendite effettuate nella città e nel periodo di riferimento. Assume solamente valori interi positivi

- **Volume:**

variabile quantitativa continua. Rappresenta il numero totale di vendite in milioni di dollari. Assume valori positivi non solamente interi.

- **Median_price:**

variabile quantitativa continua, definisce il prezzo mediano di vendita in dollari, assume valori positivi reali.

- **Month inventory:**

variabile quantitativa continua. La quantità di tempo necessaria per vendere tutte la inserzioni correnti al ritmo attuale delle vendite, espresso in mesi, assume valori positivi reali.

- **Listings:**

il numero totale degli annunci attivi una quantità intera positiva, pertanto una variabile quantitativa discreta

Sono stati calcolati i principali indici di posizione per le variabili sales, volume, median_price e listings utilizzando la funzione apply()

```
round(apply(data[, 4:7], 2, function(x) c(massimo = max(x),
        minimo = min(x), media = mean(x), mediana = median(x),dev.std = sd(x))))
```

la funzione apply() applica le funzioni max, min, mean, median e deviazione standard alle colonne 4,5,6,7 del dataset e restituisce la tabella in figura dove sono riportati minimo, massimo, media e mediana delle variabili sales, volume, median_price e listings

31	sales	volume	median_price	listings
massimo	423	84	180000	3296
minimo	79	8	73800	743
media	192	31	132665	1738
mediana	176	27	134500	1618
dev.std	80	17	22662	753

per la variabile month_inventory é stata creata una distribuzione di frequenza

	ni_month_inventory	fi_month_inventory	Ni_month_inventory	Fi_month_inventory
(3,5]	15	0.06	15	0.06
(5,7]	15	0.06	30	0.12
(7,9]	97	0.40	127	0.53
(9,11]	53	0.22	180	0.75
(11,13]	48	0.20	228	0.95
(13,15]	12	0.05	240	1.00

dalla seguente distribuzione si può notare che la terza classe (7,9] é sia la classe modale che mediana.

La variabile volume è stata divisa in classi e per essa é stata costruita la relativa distribuzione di frequenze

	ni_vol	fi_vol	Ni_vol	Fi_vol
(8,15]	41	0.17	41	0.17
(15,22]	45	0.19	86	0.36
(22,29]	44	0.18	130	0.54
(29,36]	35	0.15	165	0.69
(36,43]	24	0.10	189	0.79
(43,50]	15	0.06	204	0.85
(50,57]	15	0.06	219	0.91
(57,64]	9	0.04	228	0.95
(64,71]	7	0.03	235	0.98
(71,78]	3	0.01	238	0.99
(78,84]	2	0.01	240	1.00

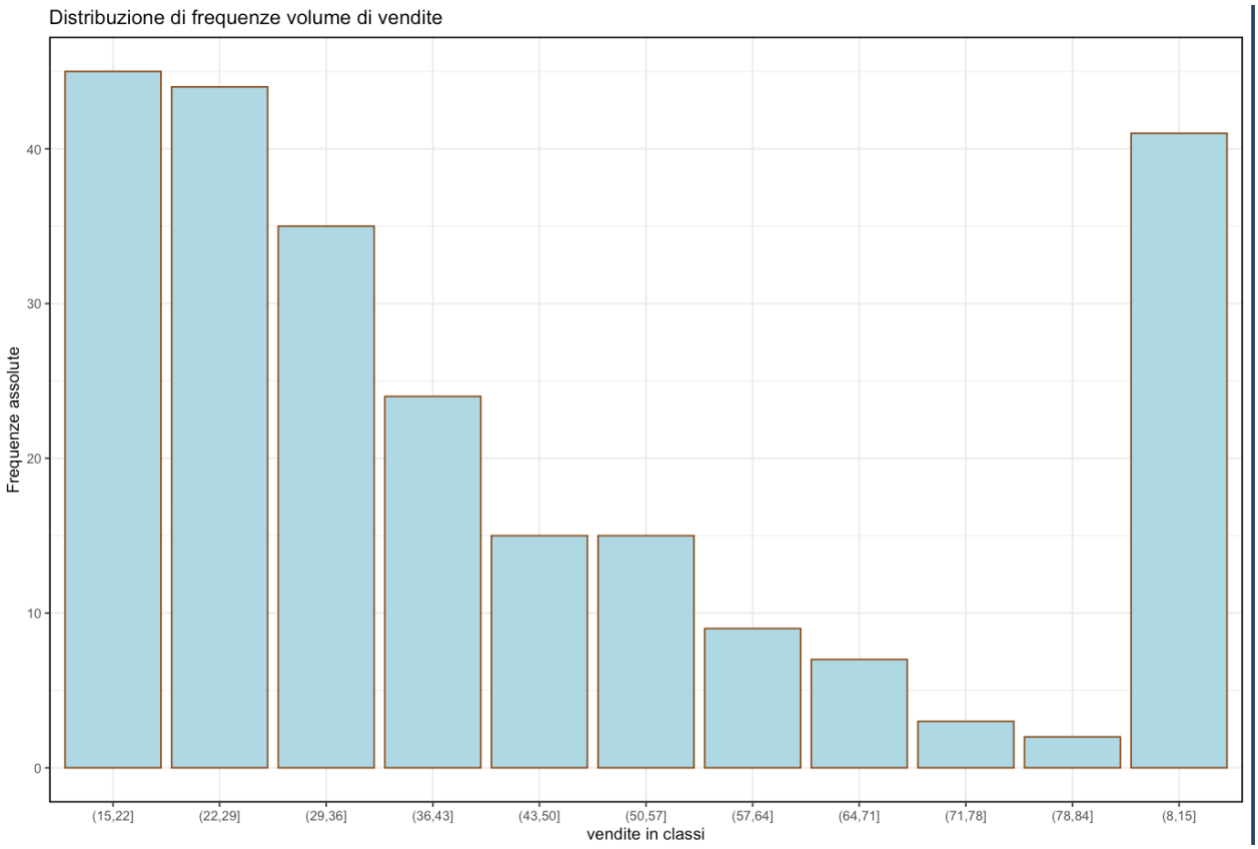
in questa distribuzione la seconda classe (15, 22] é la classe modale, mentre la terza (22.29] é la classe mediana.

```
gini.index <- function(x){
  ni = table(x)
  fi = ni/length(x)
```

```
  fi2 = fi^2
  J = length(table(x))

  gini = 1-sum(fi2)
  gini.norm = gini/((J-1)/J)

  return(gini.norm)
}

Gini_vol = gini.index(distr_freq_vol)
```

Nello script consegnato é stata definita una funzione gini.index che calcola l’indice di eterogeneità di Gini, questa funzione viene chiamata due volte nello script, nel calcolo dell’indice per la distribuzione di frequenze di volume, che risulta essere 0.3125287 e per il calcolo dell’indice della variabile city, che, come si poteva notare usando la funzione table() risulta essere 1, ovvero massima eterogeneità

Per la distribuzione di frequenze viene costruito il relativo grafico a barre che ci mostra visivamente la tabella delle distribuzioni sopra mostrata. L’immagine evidenzia come la classe (15,22] sia la classe modale



Per stabilire la variabrilitá delle singole variabili si sarebbe potuto applicare il coefficiente di varianza a tutte le variabili utilizzando la seguente funzione costruita per poi confrontarne i risultati.

```
CV <-function(x){
  return( sd(x)/mean(x) * 100 )
}
```

R ci mette a disposizione la funzione apply() con cui possiamo applicare la funzione cv costruita con la formula del coefficiente di varianza alle colonne 4,5,6,7,8.

Il risultato ottenuto viene passato ad altre due funzioni, la funzione names, che restituisce il nome delle variabili di cui abbiamo calcolato il coefficiente di varianza e max che di quei nomi restituisce solo quello della variabile con variabilità maggiore

```
variabile_var = max(names(apply(data[,4:8], 2, function(x) CV(x))))
```

Lo stesso approccio é stato utilizzato per il calcolo della variabile più asimmetrica, utilizzando in questo caso la funzione skewness messa a disposizione dal pacchetto “moments” di r

```
variabile_asim = max(names(apply(data[,4:8], 2, function(x) skewness(x))))
```

in entrambi i casi risulta che sia la variabile maggiormente asimmetrica che con variabilità maggiore é volume

Per quanto concerne le probabilità richieste, é stato utilizzata la definizione classica di probabilità, casi favorevoli/casi possibili:

- Beaumont:

la probabilità di trovare la città di Beaumont presa una riga a caso nel dataset é di 60/240, ovvero 1/4 = 0.25 o, in percentuale, del 25%

- luglio

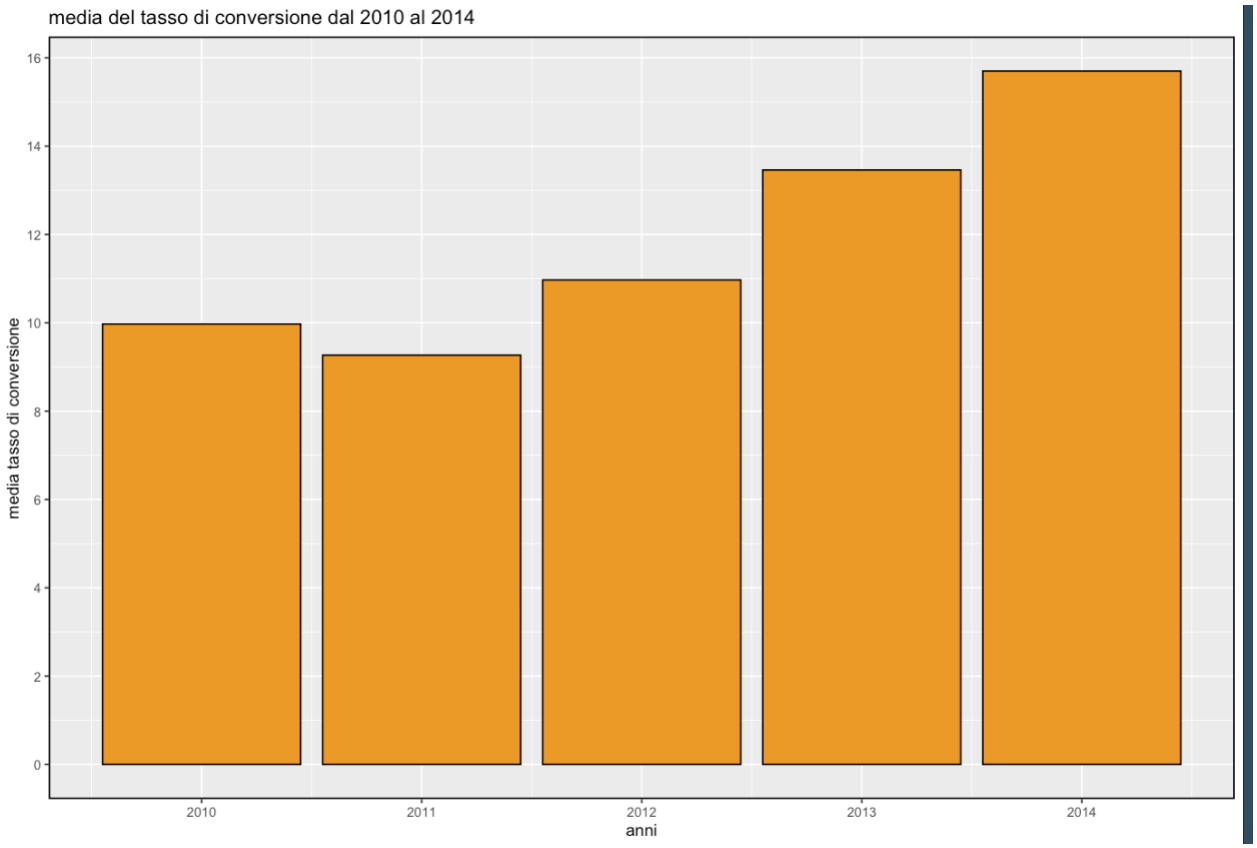
la probabilità che una riga estratta casualmente dal dataset riporti il mese di luglio é di 20/240, ovvero 1/12 = 0.083 o, in percentuale, del 8.33%

- dicembre 2012

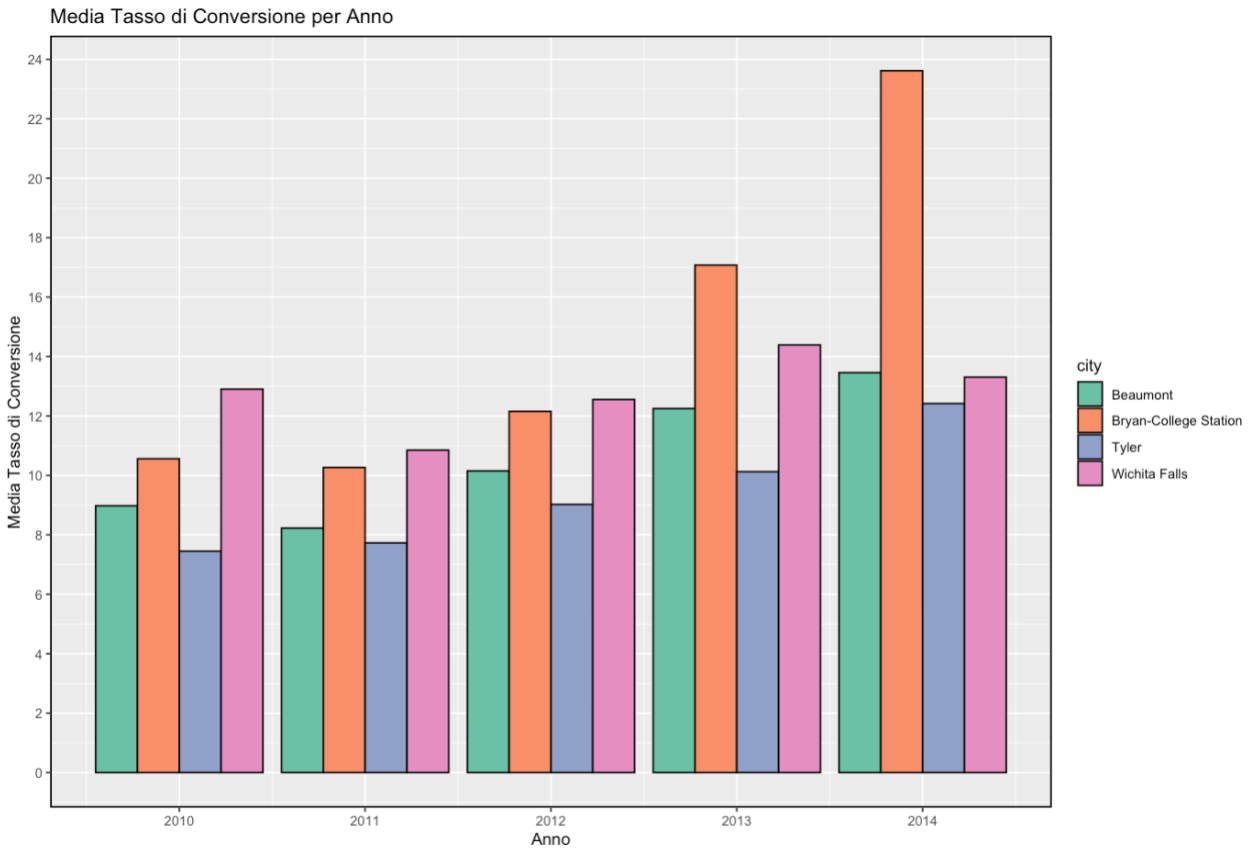
la probabilità che una riga estratta casualmente dal dataset riporti dicembre 2012 é di 4/240 ovvero 1/60 = 0.016 o, in percentuale, del 1.67%

Un’idea dell’efficacia degli annunci di vendita ci viene fornita dal tasso di conversione, espresso in percentuale. Il tasso di conversione é il rapporto tre le vendite (sales) e gli annunci (listings); per ogni n vendite ci vogliono m listings.

Dal grafico sottostante possiamo notare come la media del tasso di conversione per anno subisca un incremento costante partendo da un 10% ottenendo un incremento di 5 punti percentuali nel corso di 5 anni, solamente dal 2010 al 2011 subisce un decremento del 0.7 % risalendo tuttavia l’anno successivo ed ottenendo un incremento costante per i successivi tre anni.



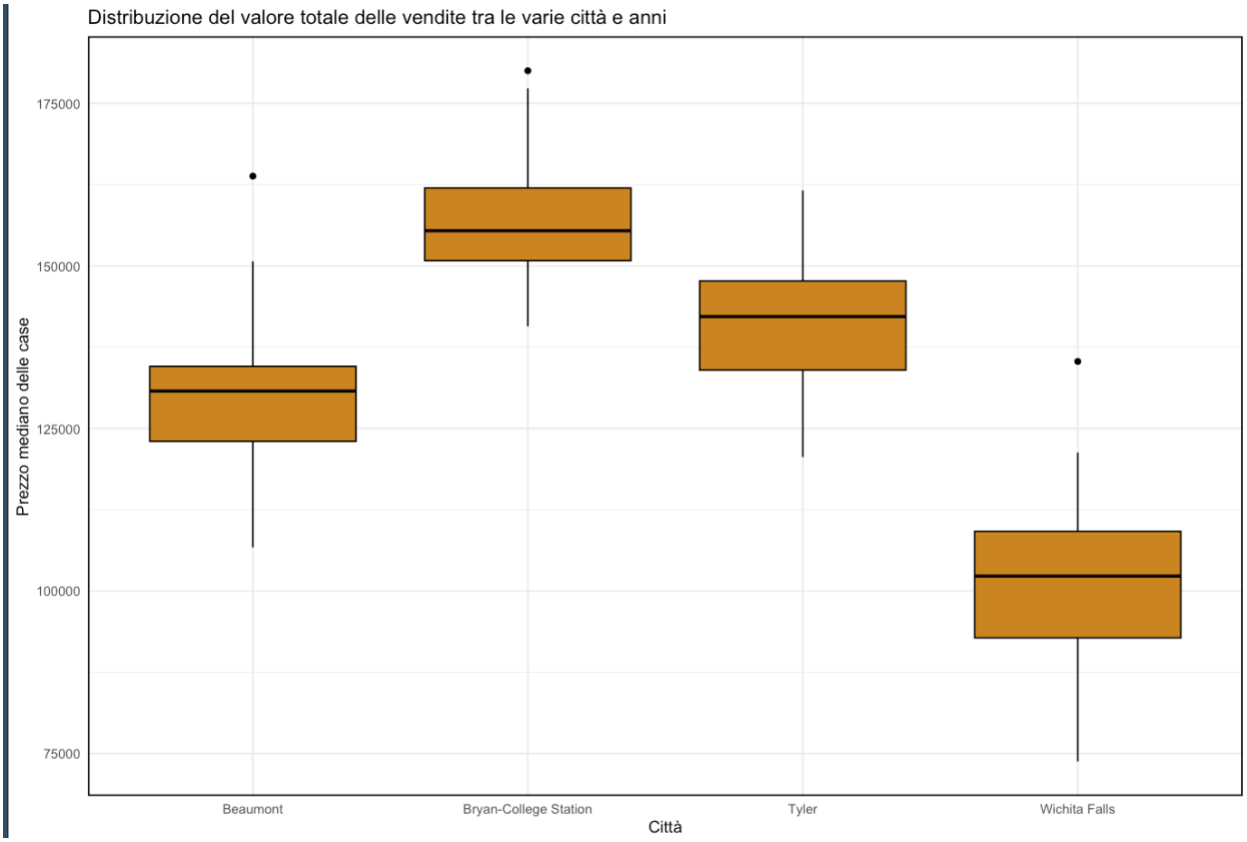
Tuttavia, inserendo anche le città nell'analisi, si nota come Tyler abbia subito un incremento costante negli anni mentre le altre tre hanno subito un decremento nel 2011 per poi avere un rendimento costante nei successivi tre anni, ad esclusione di Wichita falls che nel 2014, anno del miglior tasso di conversione delle altre città, ha subito un decremento. Bryan-college station risulta essere quella con l'incremento maggiore arrivando, nel 2014 ad avere un tasso di conversione del 23.62%



Dal confronto del pezzo mediano per città emerge che nella distribuzione dati di Tyler non sono presenti outliars mentre esistono per le altre 3 città sono presenti outliars superiori.

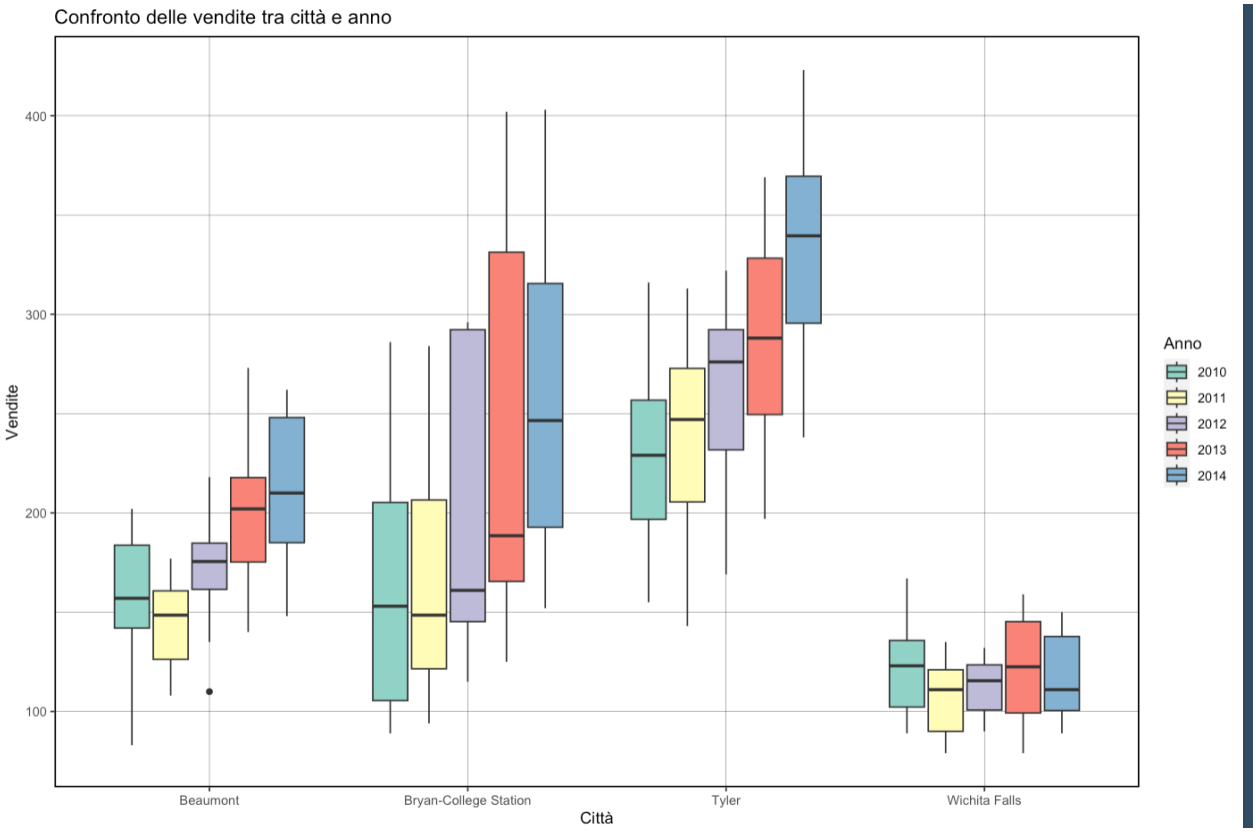
Confrontando gli IQR delle quattro città si nota come i dati raccolti siano più compatti intorno alla mediana in Beaumont e Bryan College station, mentre risultano più dispersi a Wichita Falls che presenta inoltre i prezzi mediani minori comparati alle altre località. I prezzi mediani maggiori sono quelli di Bryan college station.

1	Beaumont	11525
2	Bryan-College Station	11175
3	Tyler	13700
4	Wichita Falls	16375



Analizzando, attraverso dei boxplot il valore totale delle vendite per città negli anni, emerge che Bryan-College station ha una maggior dispersione dei dati attorno alla mediana (IQR) nei singoli anni rispetto Beaumont, Tyler e Wichita Falls, mentre proprio quest'ultima ha una distribuzione attorno alla mediana molto più compatta oltre che la città con il valore vendite minore ,sia anno per anno che nella totalità del periodo oggetto di studio.

Beaumont nel 2012 presenta outliars sia superiori che inferiori, segno del fatto che ci sono valori sia sopra la media che sotto che potrebbero influenzare la distribuzione. Sempre Beaumont ha dei whisker molto corti, ad eccezione del 2013, questo sta ad indicare che la maggior parte dei dati é compresa nel range interquartile, anche Wichita Falls presenta la stessa tendenza. Bryan college station, al contrario, presenta dei whiskers molto lunghi, che pur rimanendo nella proporzione di 1,5*IQR indicano una massiccia presenza dei dati al di fuori dello spazio Q3-Q1. Questa tendenza non si evidenzia solamente nel 2012 dove i whiskers sono minori.



IL grafico a barre sovrapposto sottostante mette a confronto le vendite nelle varie città tenendo in considerazione anche l'anno e il mese di vendita.

Nel 2010 aprile e maggio sono stati i mesi con maggiori vendite. Nel successivo anno il picco viene raggiunto da maggio e giugno. Nel 2012 ad agosto avviene il maggior numero di vendite mentre nel 2013 a luglio, infine, nel 2014, anno del miglior rendimento, nuovamente a giugno.

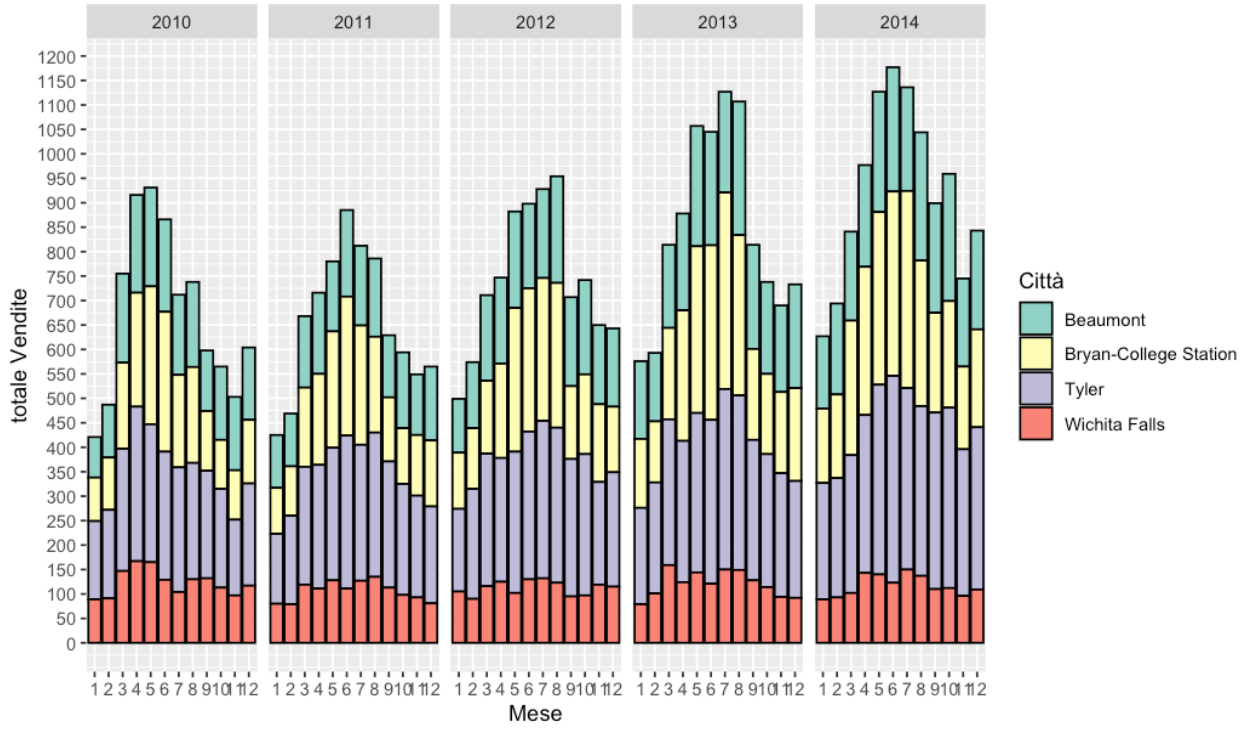
Nel 2010 l’ammontare del totale delle vendite equivale a 8096, con Tyler che supera le altre città per un totale di 2730 vendite, Bryan college station con 2011, Beaumont 1874 e Wichita Falls con 1481.

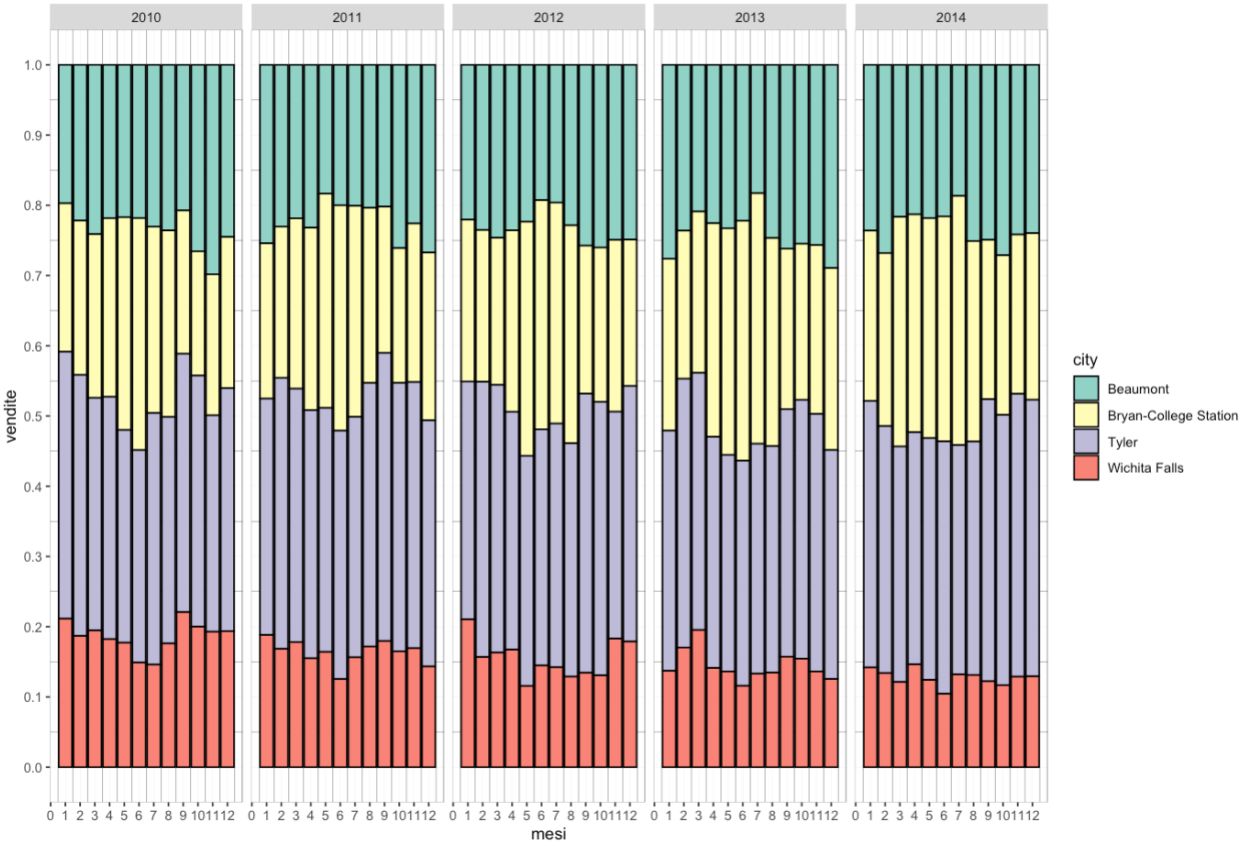
Nel 2011 subisce un calo per un totale di 7878 vendite, Tyler, tuttavia migliora i numeri dell’anno precedente con 2866 vendite, calo minimo per Bryan college station che passa a 2009 mentre più importante é la differenza di Beaumont rispetto il precedente anno con 1728 e soprattutto di Wichita Falls scende a 1275.

Nel 2012 risalgono le vendite complessive, per un totale di 8935 vendite. Tyler incrementa le vendite del 7,32%, Bryan College del 17%, Beaumont del 19% e Wichita Falls di circa il 6%.

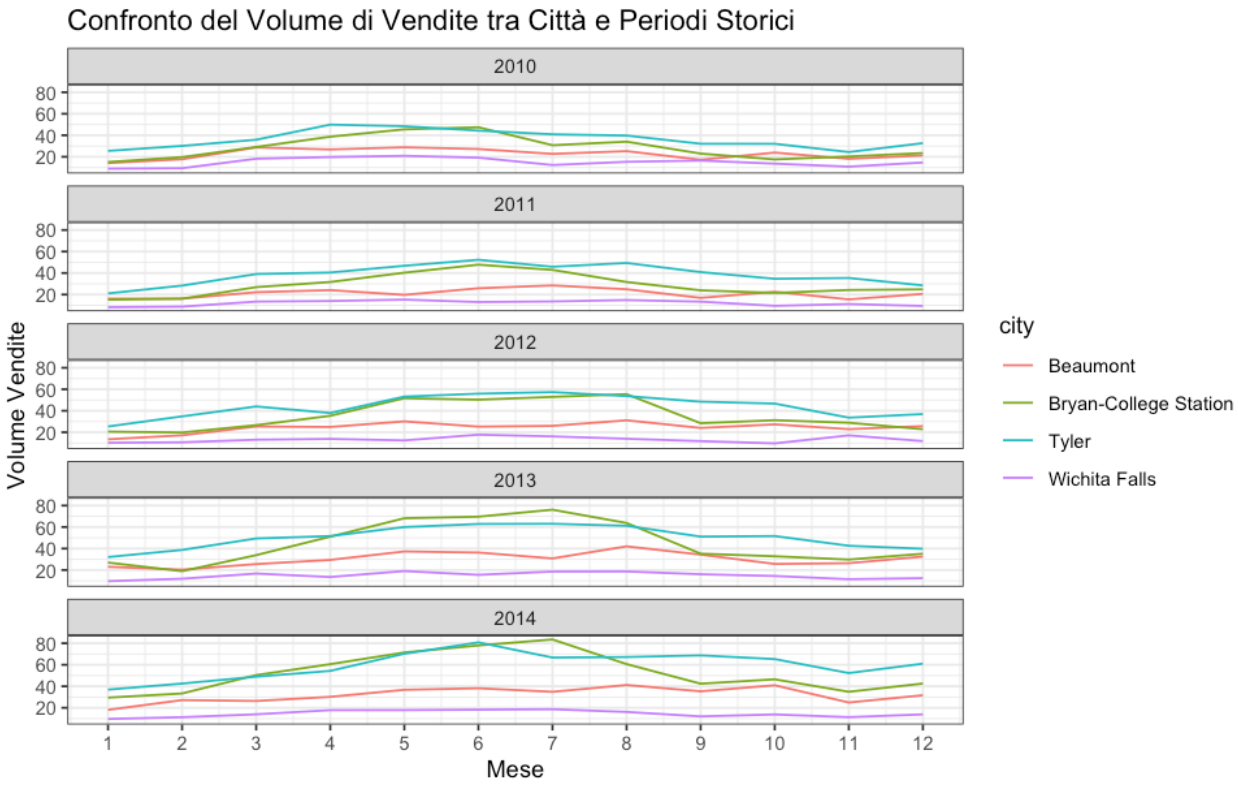
Ben più importante la crescita nel 2013, con un incremento complessivo del 13,8%, Tyler arriva a 3449 vendite, Bryan College station a 2854, Beaumont a 2414 e Wichita Falls a 1455.

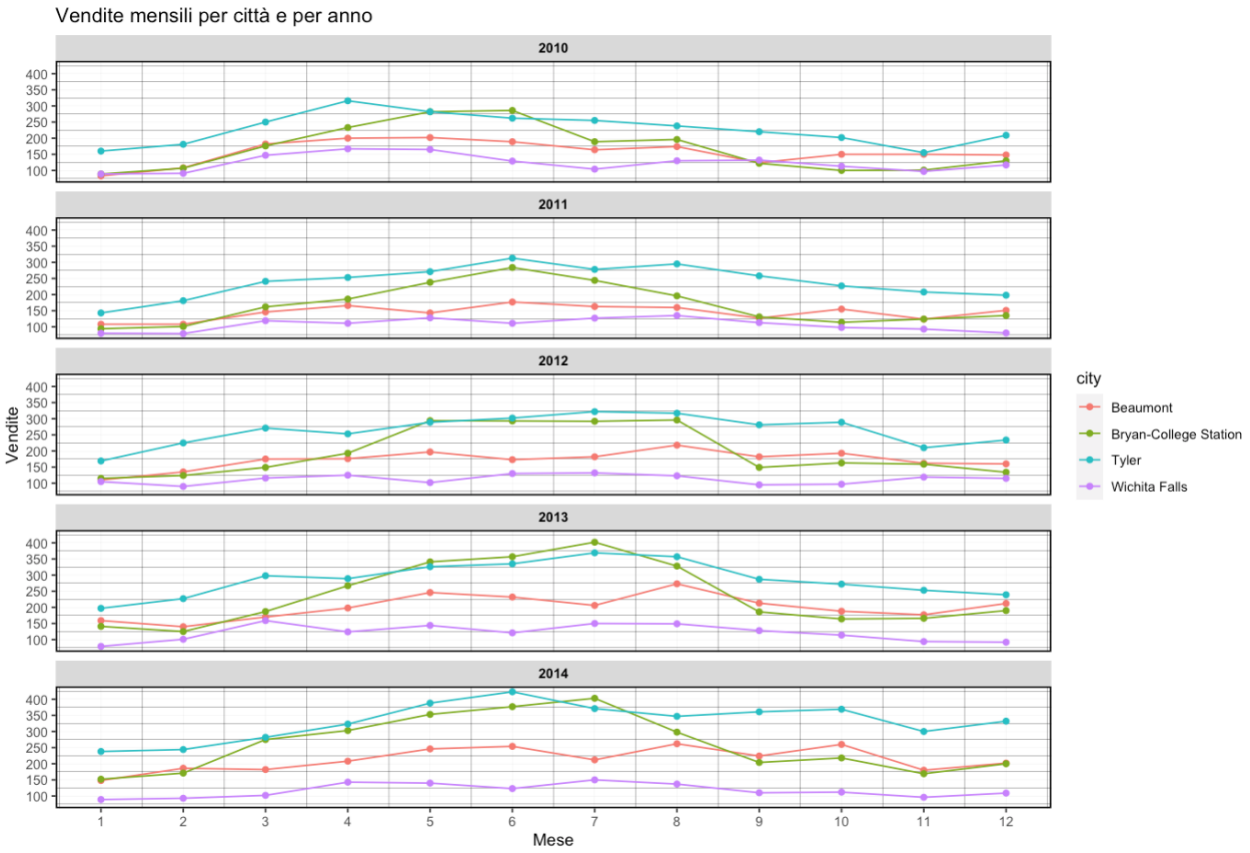
Il 2014 é il miglior anno, 11069 vendite, con un incremento di circa il 9% rispetto al precedente anno e del 36,70 % rispetto al 2010. A Tyler ci sono state 3978 vendite, a Bryan college station 3123, a Beaumont 2564 e a Wichita Falls 1404, unica in calo. Complessivamente dal 2010 al 2014, l’unica ad essere calata é stata Wichita Falls , con un decremento di oltre il 5%. Tyler ha incrementato le vendite del 45%, Beaumont del 36,84 % e Bryan College station ha avuto il miglior incremento pari al 55,24%


2010
2011
2012
2013
2014
totale Vendite
Mese
Città
Beaumont
Bryan-College Station
Tyler
Wichita Falls


2010
2011
2012
2013
2014
vendite
mesi
city
Beaumont
Bryan-College Station
Tyler
Wichita Falls

La line chart qui sotto ci consente di confrontare l’andamento dei prezzi di vendita in uno stesso anno e in anni diversi nei vari mesi per le diverse città. Esso conferma i picchi di vendita nei rispettivi mesi evidenziati dal grafico a barre sovrapposte, e ci mostra come, tendenzialmente, i picchi nei volumi di vendita avvengano nella parte centrale dell’anno e non nei 3 mesi iniziali o finali dello stesso.




Vendite mensili per città e per anno
2010
2011
2012
2013
2014
Vendite
Mese
city
Beaumont
Bryan-College Station
Tyler
Wichita Falls