Progetto Analisi esplorativa del mercato immobiliare del Texas

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

City → variabile qualitativa su scala nominale
Year → variabile qualitativa su scala ordinale
Month → variabile qualitativa su scala nominale
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.**
4. **Qual è la variabile con variabilità più elevata? Come ci sei arrivato? E quale quella più asimmetrica?**

Tabella 1. Indici di posizione

VARIABLE	MEAN	MEDIAN	MAX	MIN	1stQ	3rdQ
sales	192.29167	175.5	423	79	127	247
Volume [M$]	31.00519	27.0625	83.547	8.166	17.6595	40.893
median_price [$]	132665.41667	134500	180000	73800	117300	150050
listings	1738.02083	1618.5	3296	743	1026.5	2056
months_inventory	9.1925	8.95	14.9	3.4	7.8	10.95

Per nessuna delle variabili, il valore massimo o il valore minimo coincide (o assume un valore strettamente vicino) con quello del 1° o del 3° quantile, a significare che gli outliers sono presenti, ma non in quantità (o di valore) significativo da rendere gli indici non veritieri. Infatti, per ogni variabile, il valore medio risulta essere simile al valore mediano, indice statistico robusto, e di conseguenza anche rientrante all'interno del range interquartile (Tab. 1). In particolare, osservando il coefficiente di variazione (Tab. 2), si nota come la variabile con minor variabilità risulta essere il prezzo mediano, oscillante solo del 17%, seguita immediatamente dai mesi necessari per vendere tutte le inserzioni correnti al ritmo delle vendite (*months_inventory*, 25%).

La variabilità maggiore si osserva nel volume totale delle vendite: circa 53%. La deviazione standard (e di conseguenza anche la varianza) corrispondente a tale variabile, risulta avere un valore molto elevato, che, se sommato o sottratto al valore medio, causa un aumento o diminuzione molto significativi e quindi, un andamento generale molto variabile, come dimostra il CV (Tab. 2). Si deduce quindi che osservare il volume in un arco di tempo limitato, non sarebbe opportuno, in quanto comporterebbe la registrazione di valori non generalizzabili.

Tabella 2. Indici di variabilità

VARIABLE	IQR	Variance	SD	CV [%]
sales	120	6344.3	79.65	41.42
volume	23.23	277.27	16.65	53.71
median_price [$]	32750	513572983	22662.15	17.08
listings	1029.5	566568.97	752.71	43.31
months_inventory	3.15	5.31	2.30	25.06

*Tabella 3. Distribuzione di frequenze della variabile **month***

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

*Tabella 4. Distribuzione di frequenze della variabile **year***

ANNO	2010	2011	2012	2013	2014
FREQUENZA	48	48	48	48	48

*Tabella 5. Distribuzione di frequenze della variabile **city***

CITTA’	Beaumont	Bryan-College Station	Tyler	Wichita Falls
FREQUENZA	60	60	60	60

Osservando la distribuzione di frequenze delle variabili qualitative del dataset, notiamo come esse siano tutte monomodali, ovvero presentino sempre la stessa frequenza (Tab. 4, 5, 6). Ci aspettiamo quindi che calcolando l’indice di Gini, esso sia uguale a 1. Inoltre, ciò permette di fare un’analisi più accurata delle variabili quantitative, in quanto misurazioni effettuate nelle medesime condizioni e frequenze appunto di mesi, anni e città, rendendo i risultati sicuramente confrontabili e non soggetti a differenze dovute a stagionalità o influenza culturale.

Tabella 6. Indice di asimmetria di Fisher e Curtosi

VARIABILE	SKEWNESS	KURTOSIS
sales	0.718	-0.313
volume	0.884	0.177
median_price	-0.365	-0.623

listings	0.649	-0.792
months_inventory	0.041	-0.174

Osservando l'indice di asimmetria di Fisher, notiamo come esso sia positivo per tutte le variabili, ad eccezione del prezzo mediano. Il che significa che la maggior parte delle variabili presenta una distribuzione di frequenze asimmetrica positiva, con valore medio maggiore del valore mediano (come si può osservare, per controprova, dalla tabella degli indici di posizione, Tab. 1). La variabile *months_inventory* è quella per cui l'indice risulta essere il più vicino allo zero, a significare che presenta una distribuzione molto vicina a quella simmetrica. Sebbene infatti il suo valore medio sia maggiore rispetto a quello mediano, la distanza tra i due valori è minima.
Per il prezzo mediano, invece la distribuzione è asimmetrica negativa, con conseguente media minore della mediana.
In generale, la variabile più asimmetrica, ovvero con valore più discostato dallo 0, è rappresentata dal volume.
Per quanto concerne invece l'indice di curtosi, esso è negativo per tutte le variabili, ad eccezione del volume. Le distribuzioni sono perciò principalmente platicurtiche, ovvero più appiattite rispetto ad una distribuzione normale. L'unica distribuzione leptocurtica è, come anticipato, quella relativa al volume.
Mettendo a confronto tutte le variabili e i corrispettivi indici di asimmetria e curtosi, quella con forma più discostata dalla distribuzione normale è il numero di inserzioni (*listings*), che presenta entrambi gli indici molto lontani dallo zero.

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.

Gini Index della variabile *sales*: 0.998379
Gini index calcolato sulla variabile *sales* suddivisa in classi: 0.8993981

L'indice di Gini per la variabile *sales* risulta essere molto vicino a 1. Ciò significa che la distribuzione dei valori risulta essere molto vicina ad una distribuzione omogenea. Lo stesso indice calcolato sulla variabile suddivisa in classi risulta avere un valore più basso, andando ad indicare che con tale suddivisione si ottiene maggiore eterogeneità, come si può notare visivamente dal grafico sottostante (Fig. 1):

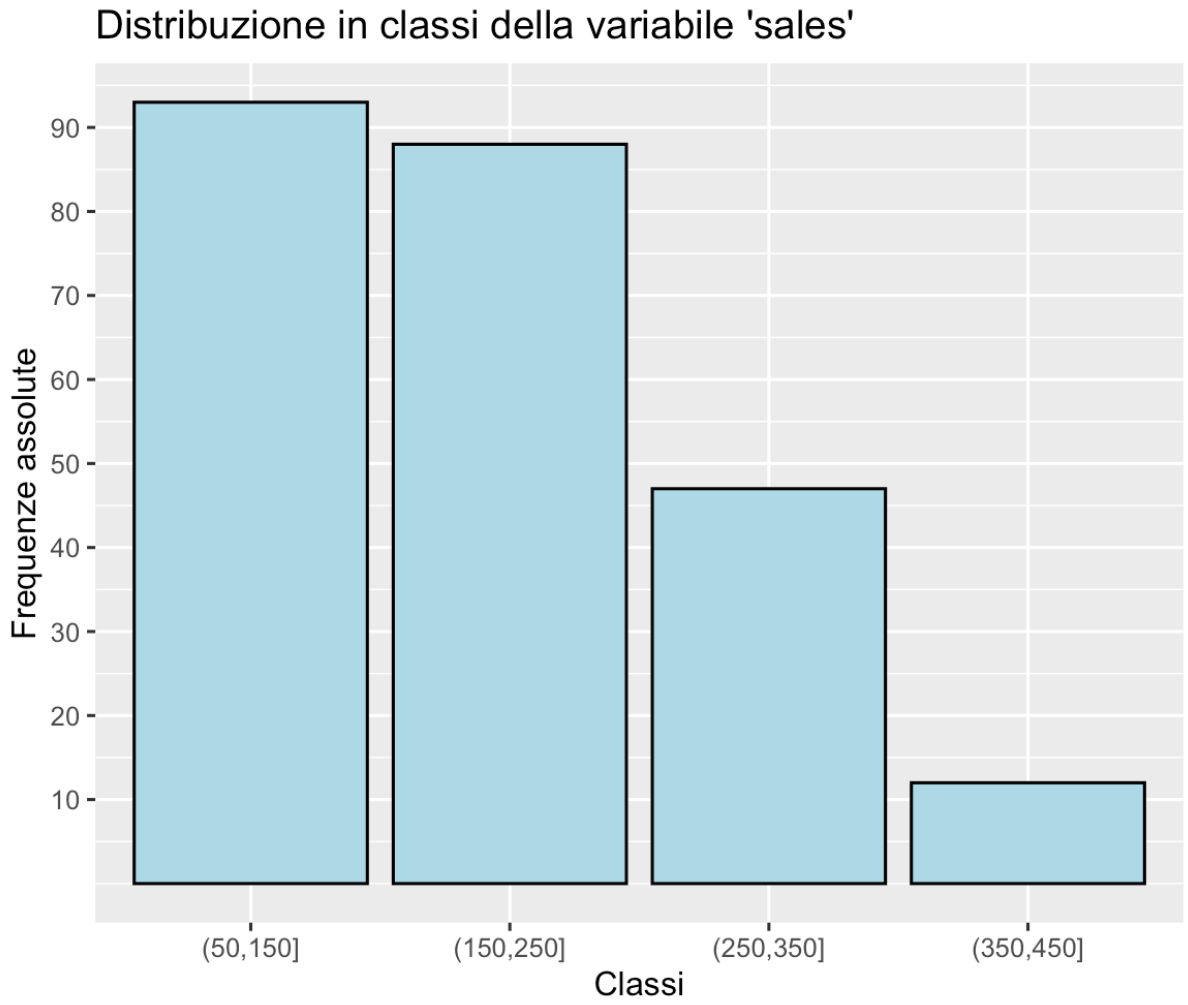

Figura 1

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

Avendo la variabile *city* una distribuzione monomodale, il suo indice di Gini è pari a 1.

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?

Tabella 7. Probabilità

Caso	Probabilità
Beaumont	25 %
Luglio	8 %
Dicembre 2012	1.2 %

Si veda il file .R per maggiori dettagli (da riga 106).

8. Esiste una colonna col prezzo mediano, creane una che indica invece il prezzo medio, utilizzando le altre variabili che hai a disposizione.
Il valore del prezzo medio può essere calcolato dividendo il volume (convertito in dollari) per il numero di vendite nel singolo mese. Si veda il file .R per i valori (da riga 121).

9. Prova a creare un'altra colonna che dia un'idea di "efficacia" degli annunci di vendita. Riesci a fare qualche considerazione?
L'efficacia degli annunci ci vendita può essere espressa come percentuale delle vendite effettuate in ogni mese. Dividendo il valore della variabile *sales* per quello della variabile *listings*, si ottiene un valore che esprime quanti annunci hanno effettivamente comportato una vendita. Maggiore è il rapporto ottenuto, maggiore è da intendersi l'efficacia degli annunci in un determinato mese. Si veda il file .R per i valori (da riga 126).

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. Puoi utilizzare il linguaggio R di base oppure essere un vero Pro con il pacchetto dplyr. Ti lascio un suggerimento in pseudocodice, oltre al cheatsheet nel materiale:

```
dati %>%
group_by(una o più variabili di raggruppamento) %>%
summarise(nomecolonna1=funzione1(variabile da sintetizzare),
          nomecolonna2=funzione2(variabile da sintetizzare))
```

Sfruttando questa notazione puoi creare anche dei grafici super!

Da qui in poi utilizza ggplot2 per creare grafici fantastici!

Ma non fermarti alla semplice soluzione del quesito, prova un po' a personalizzare i grafici utilizzando temi, colori e annotazioni, e aggiustando i vari elementi come le etichette, gli assi e la legenda.
Consiglio: Fai attenzione quando specifichi le variabili month e year tra le estetiche, potrebbe essere necessario considerarle come fattori.

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

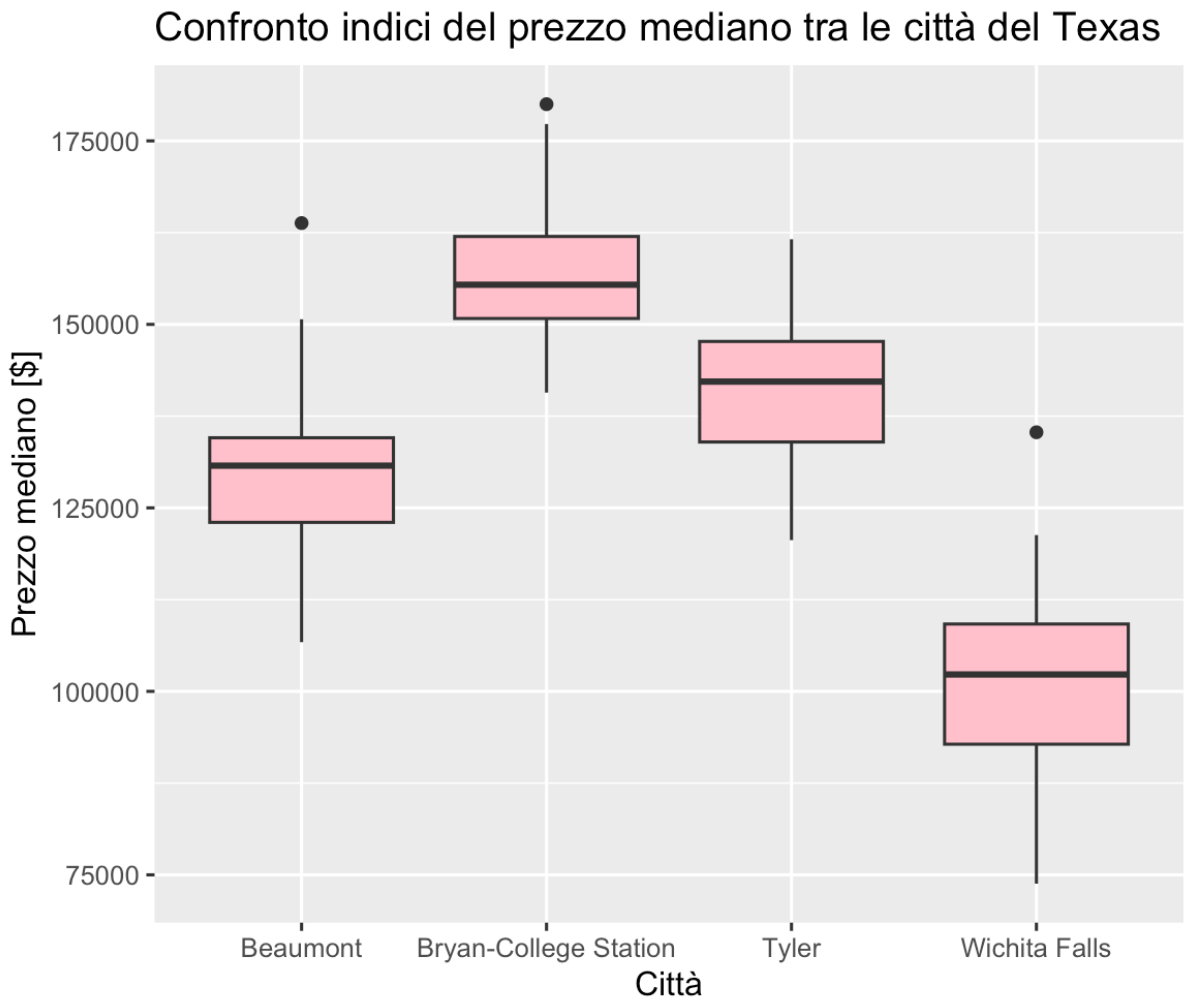

Figura 2

Tabella 8. Confronto indici di posizione e di variabilità tra le varie città del Texas

City	Mean [$]	SD [$]	CV [%]	Max [$]	Median [$]	Min [$]
Beaumont	129988	10104.99	7.77	163800	130750	106700
Bryan-College Station	157488	8852.23	5.62	180000	155400	140700
Tyler	141441	9336.53	6.60	161600	142200	120600
Wichita Falls	101743	11320.03	11.12	135300	102300	73800

Osservando i boxplot relativi alla distribuzione del prezzo mediano delle case tra le varie città del Texas, si nota immediatamente come i valori maggiori si riscontrino per Bryan-College Station e quelli minori per Wichita Falls.
Questo dato ci è suggerito non solo dal valore mediano dei singoli boxplot, ma anche dall'ampiezza e posizionamento dei range interquartili: nella città di Bryan-College Station, ad esempio, esso si colloca tra i 150000$ e i 162500$ circa, valori visivamente maggiori rispetto alle altre città. Un aspetto sicuramente comune per tutte è l'asimmetria delle distribuzioni: in nessuno dei boxplot la mediana divide il box in parti uguali, indicando quindi la presenza di una certa asimmetria. Una ulteriore considerazione può essere fatta rispetto alla presenza di outliers per tutte le città, eccetto Tyler.

2. 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?

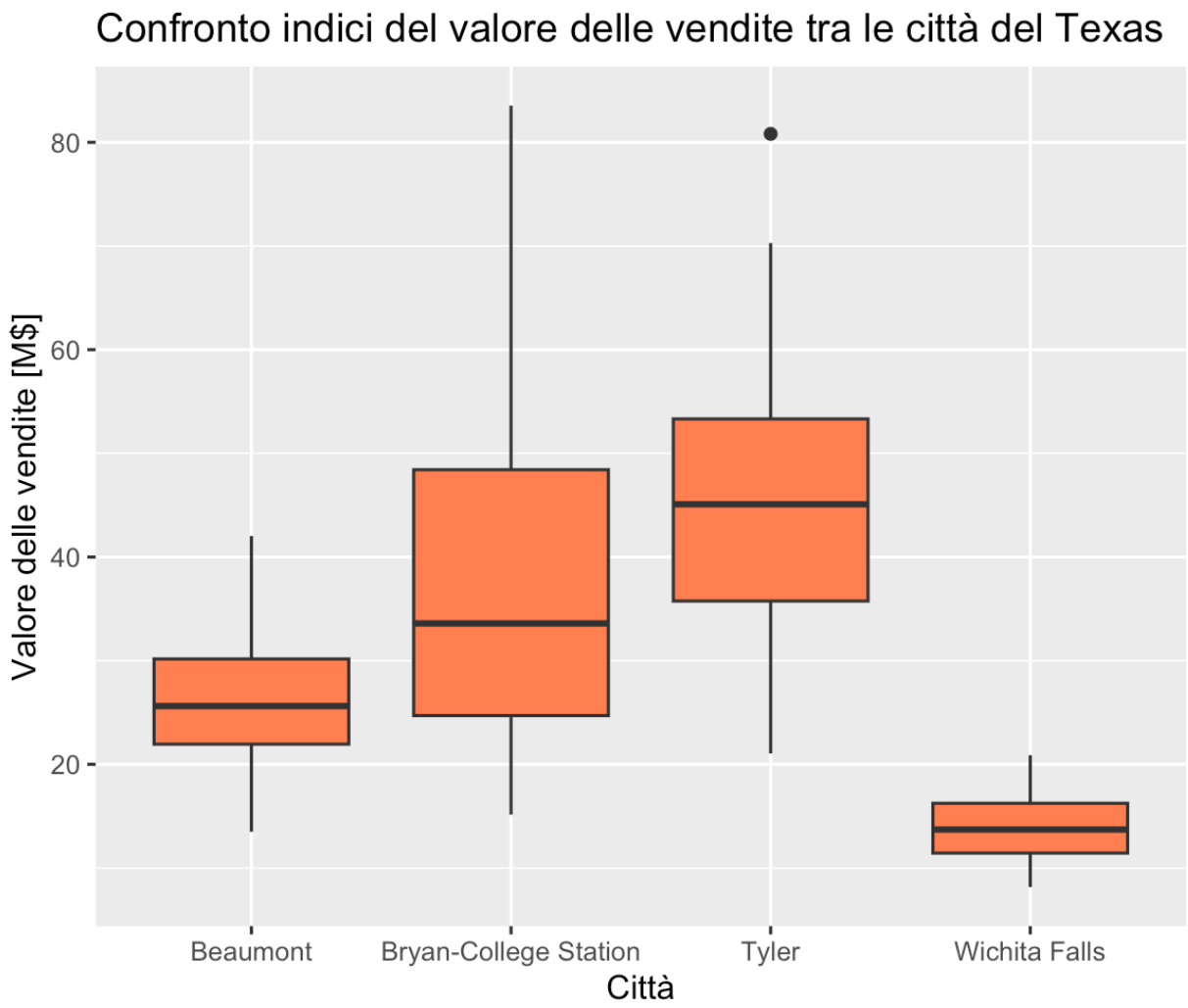

Figura 3

Andando a confrontare le distribuzioni del valore delle vendite tra le varie città, notiamo come nella città di Tyler il valore delle vendite sia tendenzialmente maggiore rispetto alle altre città (si veda il valore della mediana), sebbene parte del range interquartile si intersechi con quello di Bryan-College Station. Confrontando questi risultati con quelli riportati nel grafico precedente (Fig. 2), deduciamo che nella città di Tyler ci sia stato un maggior numero di vendite negli anni rispetto alla città di Bryan-College Station, poiché, sebbene Tyler riporti un prezzo mediano di vendita minore, allo stesso tempo il valore totale risulta maggiore. La controprova di tale deduzione è visibile dal grafico di figura 4.
La città di Wichita Falls risulta invece avere un valore totale delle vendite minore rispetto a tutte le altre città, conseguentemente al fatto che ha il minor numero di vendite e il minor prezzo mediano di vendita. Presenta invece una distribuzione molto vicina ad una distribuzione simmetrica ed un range interquartile ridotto, a significare che la maggior parte dei dati assumono valori molto vicini al valore mediano.
Le distribuzioni dei dati per le altre città sono invece più asimmetriche.

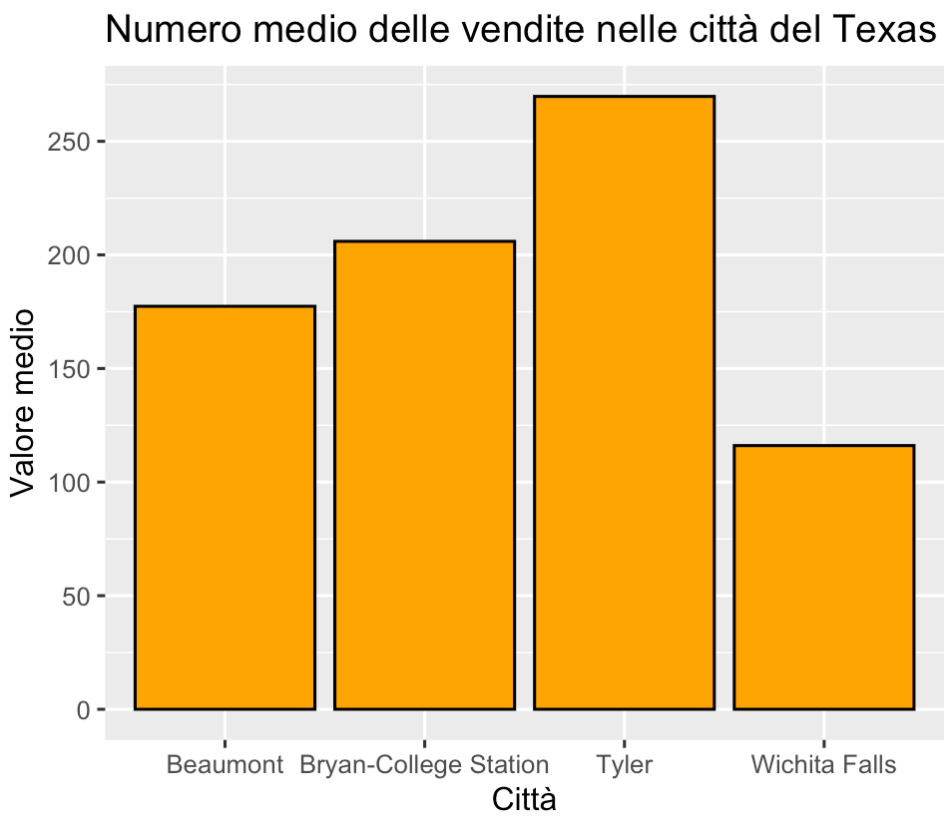

Figura 4

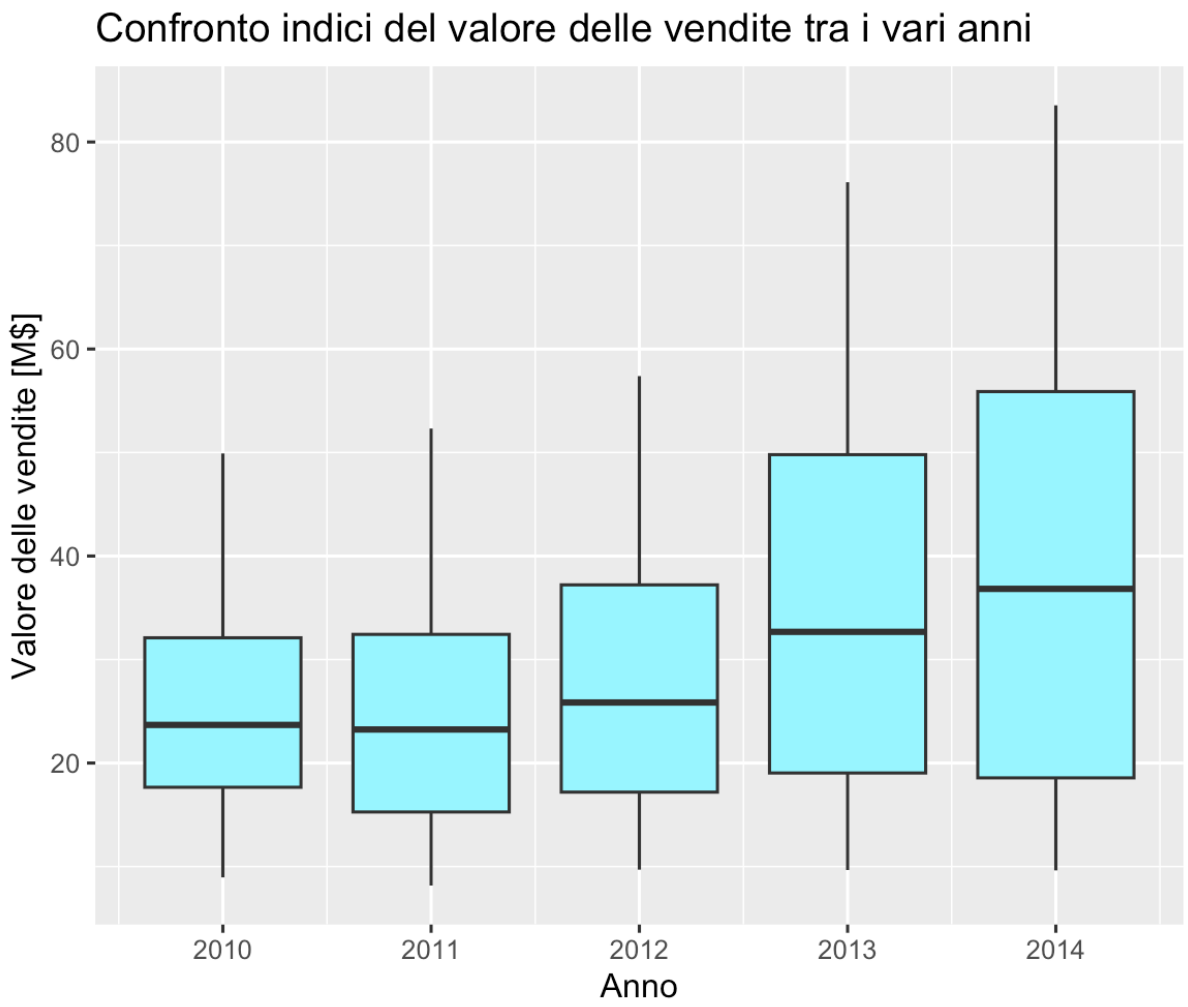

Figura 5

La prima considerazione da fare osservando i boxplot relativi al valore delle vendite negli anni (Fig. 5) riguarda l'aumento dell'ampiezza dei box, ad indicare una crescita nella variabilità della variabile in questione. In particolar modo, il minimo del valore delle vendite ha subito delle minime, ed in alcuni momenti quasi nulle, oscillazioni: la variabilità è quindi dovuta ad un aumento generale dei prezzi, che ha causato un allargamento della campana delle distribuzioni. Di conseguenza, anche il valore mediano si è spostato verso valori maggiori nel corso degli anni.

3. 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. Consiglio: Stai attento alla differenza tra geom_bar() e geom_col(). PRO LEVEL: cerca un modo intelligente per inserire ANCHE la variabile Year allo stesso blocco di codice, senza però creare accrocchi nel grafico.

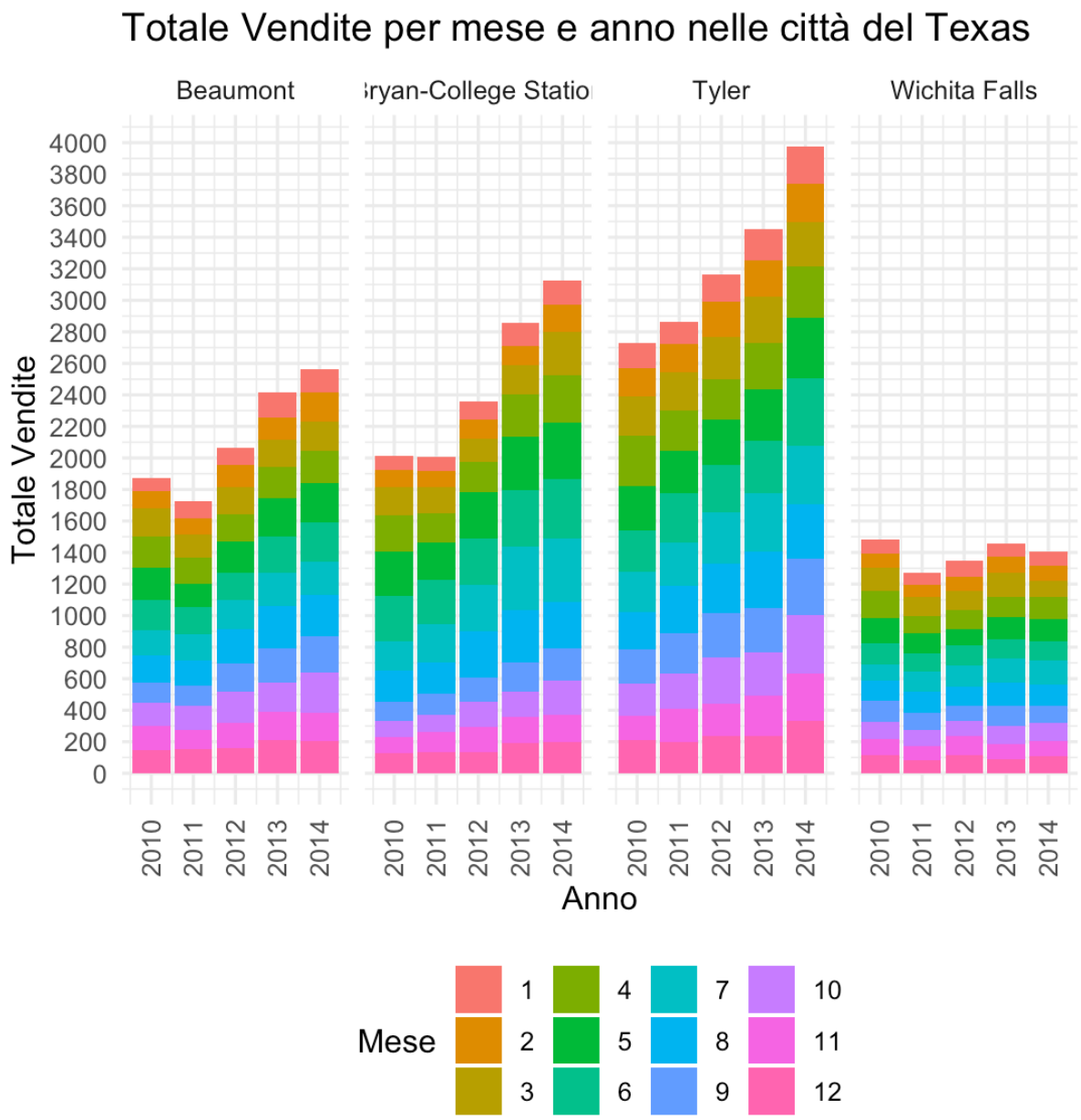

Figura 6

Come visibile dal grafico di figura 6, e già precedentemente mostrato in figura 4, la città di Tyler si rivela essere quella per cui nel corso degli anni il totale delle vendite è sempre stato maggiore rispetto alle altre città. Ad esempio, si nota come la quantità delle vendite nella città di Tyler nel 2012, è stata raggiunta dalla città di Bryan-College Station solo nel 2014. Nelle città di Beaumont e Wichita Falls invece, le vendite non hanno mai raggiunto quelle della città di Tyler, nemmeno nel corso del tempo o confrontando i valori con l'anno 2010 di Tyler. Inoltre, mentre per Tyler e Bryan-College Station l'andamento nel corso degli anni è stato crescente, la stessa considerazione non può essere fatta per Beaumont e Wichita Falls, dove l'anno 2011 ha visto in entrambi i casi un calo nelle vendite, ripetutosi per Wichita Falls anche nel 2014.
Un'analisi sulle vendite nei vari mesi può essere fatta in maniera più opportuna osservando il grafico normalizzato (Figura 7). In termini percentuali, le maggiori vendite sono state fatte a

Bryan-College Station nei mesi di maggio e giugno (5 e 6) in tutti gli anni. Qualitativamente parlando, il mese di gennaio (1) appare invece essere tendenzialmente il mese in cui in ogni città sono state fatte meno vendite.

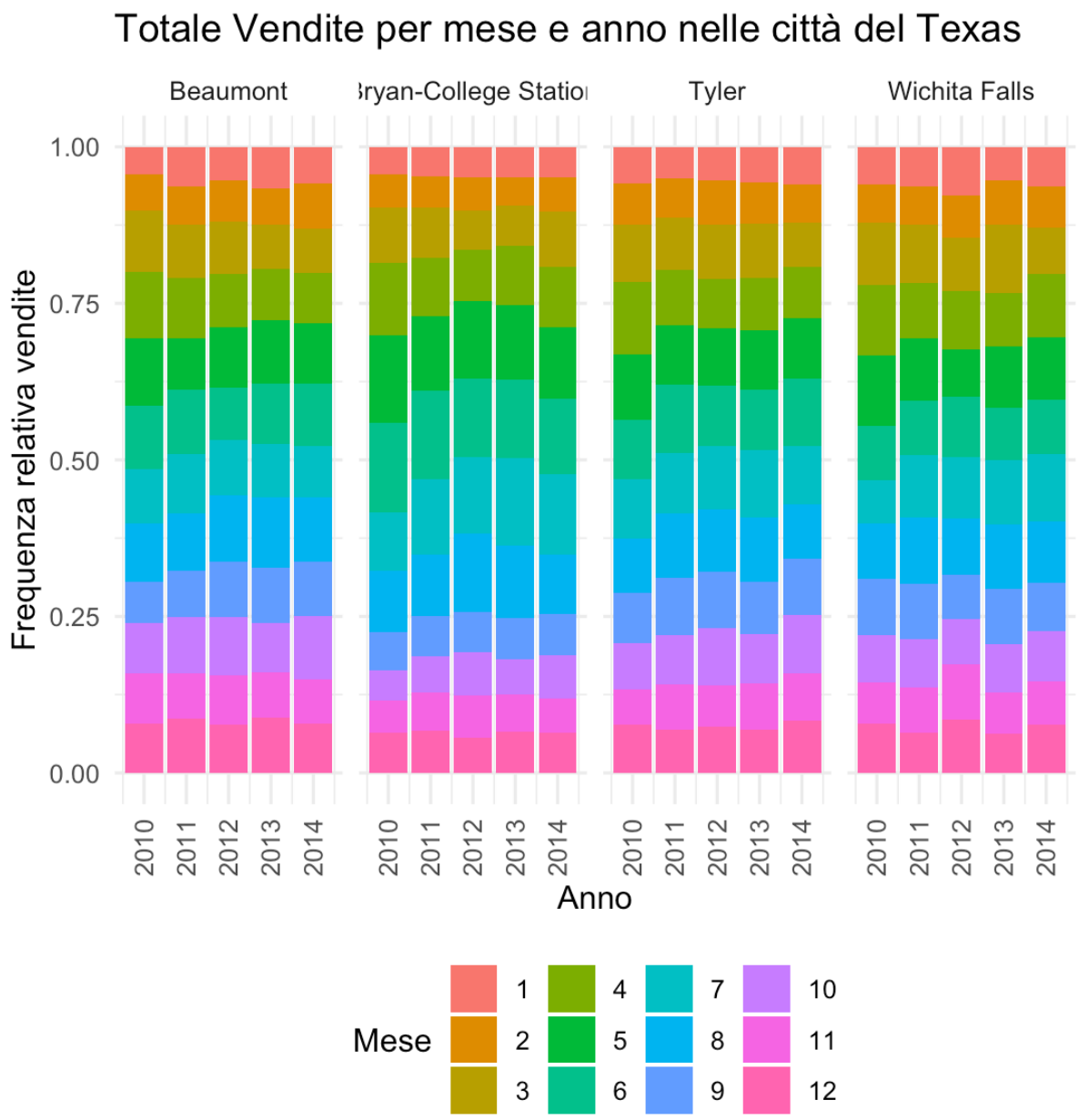

Figura 7

4. Crea un line chart di una variabile a tua scelta per fare confronti commentati fra città e periodi storici. Ti avviso che probabilmente all'inizio ti verranno fuori linee storte e poco chiare, ma non demordere. Consigli: Prova inserendo una variabile per volta. Prova a usare variabili esterne al dataset, tipo vettori creati da te appositamente. Se non riesci proprio a venirne a capo inizia lavorando su dataset ridotti, ad esempio prendendo in considerazione un solo anno o una sola città. Aiutati con il pacchetto dplyr:

dati2014 <- filter(dati, year==2014)

dati_Beaumont <- filter(dati, city==”Beaumont”)

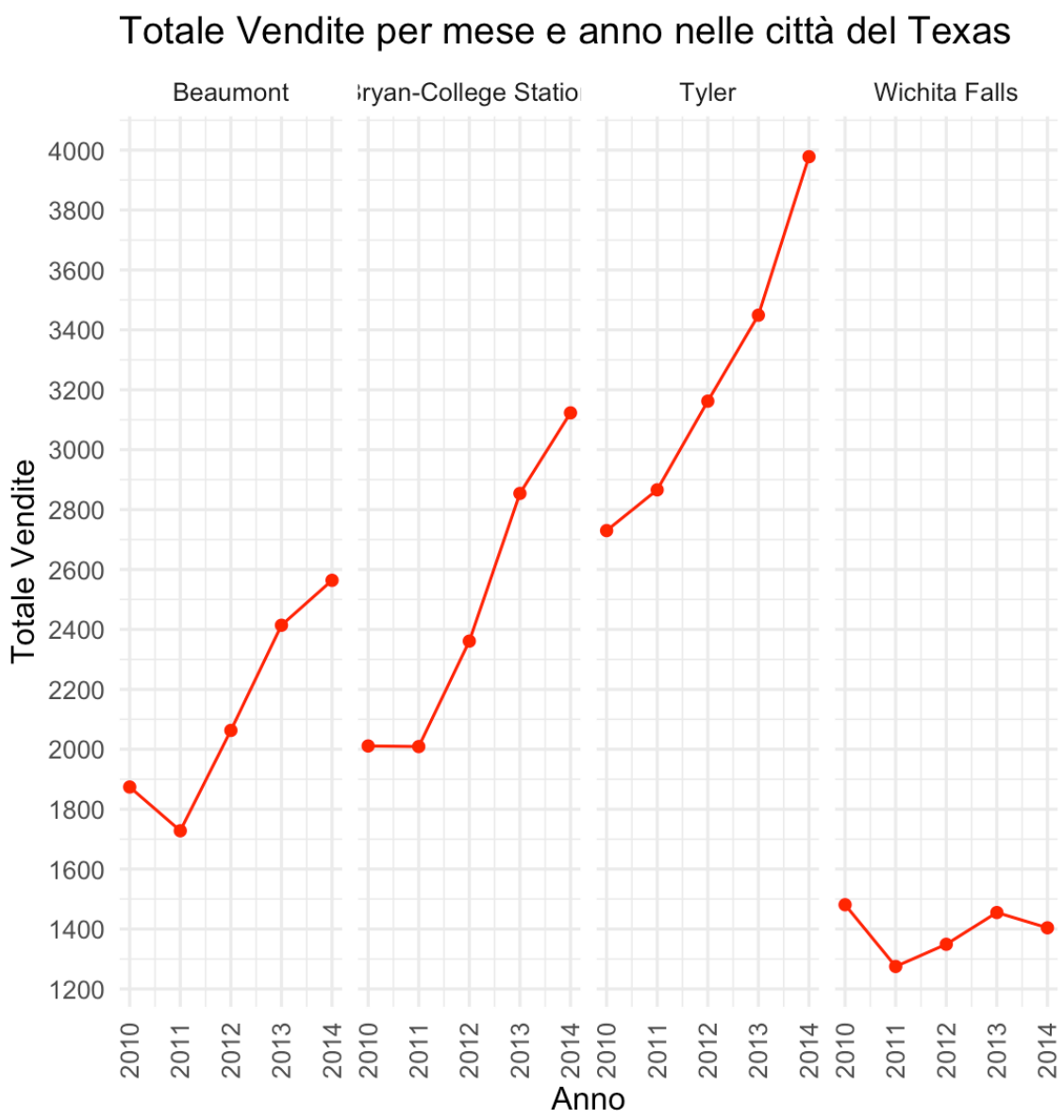

Figura 8

Rispetto agli istogrammi di figura 6, dai line chart di figura 8 è possibile ricavare più facilmente informazioni relativi all’incremento (o decremento) delle vendite tra un anno e l’altro e metterlo a confronto fra le varie città. Ad esempio, incrementi significativi delle vendite si sono registrati tra il 2012 e il 2013 per la città di Bryan-College Station e tra il 2013 e il 2014 per Tyler. Nella città di Beaumont notiamo una crescita con un andamento approssimabile a lineare tra il 2011 e il 2013, mentre nessuna apparente variazione tra il 2010 e il 2011 a Bryan-College Station (andando a rappresentare il grafico con una maggiore risoluzione o andando ad osservare I valori numerici precisi è possibile che possano essere notate delle piccole variazioni, ma sicuramente poco significative, che per un’analisi ad alto livello possono essere approssimate ad un andamento costante). Le uniche città in cui si registra un decremento nelle vendite sono Beaumont e Wichita Falls (qui più ripido) tra il 2010 e il 2011 e, solo per Wichita Falls, nuovamente tra il 2013 e il 2014.

In figura 9 sono rappresentati i line chart del numero di vendite e di annunci a confronto per ogni anno e in ogni città, che ci permette di analizzare in maniera visiva l'efficacia degli annunci. Ogni qualvolta si osserva la curva del numero di annunci descrescere e allo stesso tempo quella delle vendite corrispondenti crescere, possiamo affermare che nel tempo l'efficacia è aumentata, perchè ad un numero minore di annunci è corrisposto un numero maggiore di vendite. Questa situazione si è verificata almeno una volta in tutte le città, con particolare efficacia tra il 2012 e il 2014 per la città di Tyler, essendoci stato un decremento importante nella quantità di annunci esposti. In tutte le città è stato fatto un tentativo di aumento degli annunci nel primo anno (tra il 2010 e il 2011), situazione che si è ripetuta nuovamente solo per la città di Wichita Falls tra il 2013 e il 2014, con meno efficacia rispetto al periodo precedente.
D'altra parte, mettendo a confronto l'efficacia degli annunci tra le varie città osservando la distanza tra le curve della vendita e degli annunci, è possibile ipotizzare che gli annunci nella città di Tyler siano in generale i meno efficaci: sebbene, come analizzato in precedenza (ad esempio osservando figura 4 e 8), essa sia la città con il maggior numero di vendite rispetto alle altre, ciò è probabilmente dovuto al fatto che vengono esposti molti più annunci, molti dei quali non comportano però una conseguente vendita. A conferma di ciò è possibile osservare I valori numerici di Tabella 9.

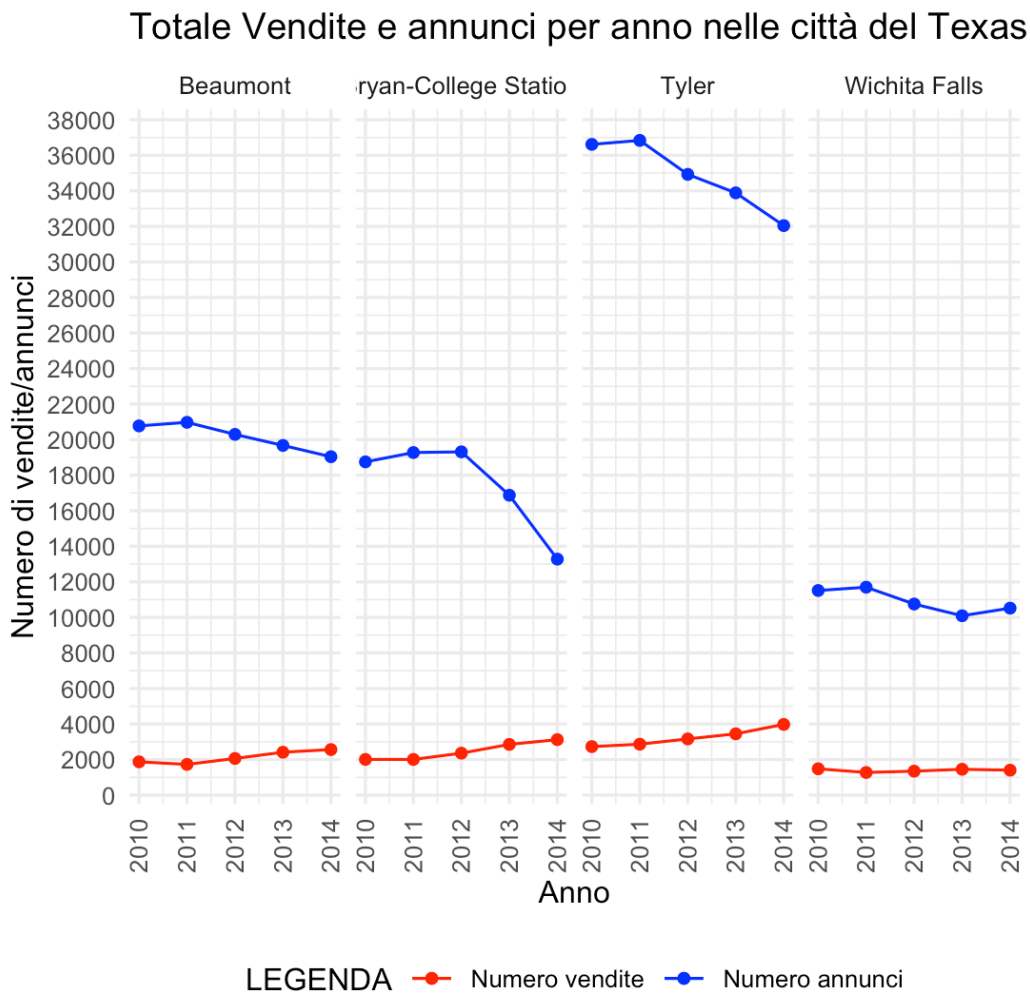

Figura 9

Tabella 9

Città	Anno	Numero annunci	Numero vendite	Efficacia [%]
Beaumont	2010	20773	1874	9.0
	2011	20975	1728	8.2
	2012	20296	2063	10.2
	2013	19675	2414	12.3
	2014	19040	2564	13.5
Bryan-College Station	2010	18749	2011	10.7
	2011	19274	2009	10.4
	2012	19314	2361	12.2
	2013	16873	2854	16.9
	2014	13278	3123	23.5
Tyler	2010	36613	2730	7.5
	2011	36836	2866	7.8
	2012	34925	3162	9.1
	2013	33885	3449	10.2
	2014	32044	3978	12.4
Wichita Falls	2010	11513	1481	12.9
	2011	11698	1275	10.9
	2012	10752	1349	12.5
	2013	10092	1455	14.4
	2014	10520	1404	13.3