# Progetto

# di

# Statistica Descrittiva

Santone Carmine

c.santone@outlook.it

## 1) Scarica il dataset realestate_textas.csv da qui e importalo con R

```
> setwd("C:/Users/csant/Desktop/ProfessionAI/Statistica Descrittiva")

> library(readr)

> texas <- read.csv("realestate_texas.csv")
```

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

*City*: variabile qualitativa nominale

*Year*: anche se espressa con numeri interi è da considerare come variabile qualitativa ordinale

*Month*: per la motivazione espressa precedentemente è una variabile qualitativa ordinale

*Sales*: variabile quantitativa

*Volume*: variabile quantitativa

*Median price*: variabile quantitativa

*Listings*: variabile quantitativa

*Months inventory*: variabile quantitativa

## 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

| | ni.city | fi.city |
|---|---|---|
| **Beaumont** | 60 | 0.25 |
| **Bryan-College Station** | 60 | 0.25 |
| **Tyler** | 60 | 0.25 |
| **Wichita Falls** | 60 | 0.25 |

| | ni.year | fi.year | Ni.year | Fi.year |
|---|---|---|---|---|
| **2010** | 48 | 0.2 | 48 | 0.2 |
| **2011** | 48 | 0.2 | 96 | 0.4 |
| **2012** | 48 | 0.2 | 144 | 0.6 |
| **2013** | 48 | 0.2 | 192 | 0.8 |
| **2014** | 48 | 0.2 | 240 | 1.0 |

| | ni.month | fi.month | Ni.month | Fi.month |
|---|---|---|---|---|
| **1** | 20 | 0.08333333 | 20 | 0.08333333 |
| **2** | 20 | 0.08333333 | 40 | 0.16666667 |
| **3** | 20 | 0.08333333 | 60 | 0.25000000 |
| **4** | 20 | 0.08333333 | 80 | 0.33333333 |
| **5** | 20 | 0.08333333 | 100 | 0.41666667 |
| **6** | 20 | 0.08333333 | 120 | 0.50000000 |

| | | | | |
|---|---|---|---|---|
| 7 | 20 | 0.08333333 | 140 | 0.58333333 |
| 8 | 20 | 0.08333333 | 160 | 0.66666667 |
| 9 | 20 | 0.08333333 | 180 | 0.75000000 |
| 10 | 20 | 0.08333333 | 200 | 0.83333333 |
| 11 | 20 | 0.08333333 | 220 | 0.91666667 |
| 12 | 20 | 0.08333333 | 240 | 1.00000000 |

Per le variabili “city”, “years” e “month”, essendo qualitative, ho creato solo le distribuzioni di frequenza. Come si può vedere, all’interno del dataset, il numero di osservazioni per ogni città è identico e sono stati raccolti dati per ogni mese dal 2010 al 2014.

| | Mean | Mode | Median | Min | Max | Range | IQR | Variance | SD | CV | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| sales | 192.29167 | 124.000 | 175.5000 | 79.000 | 423.000 | 344.000 | 120.0000 | 6.344300e+03 | 79.651111 | 0.4142203 | 0.71810402 | -0.3131764 |
| volume | 31.00519 | 14.003 | 27.0625 | 8.166 | 83.547 | 75.381 | 23.2335 | 2.772707e+02 | 16.651447 | 0.5370536 | 0.88474203 | 0.1769870 |
| median_price | 132665.41667 | 130000.000 | 134500.0000 | 73800.000 | 180000.000 | 106200.000 | 32750.0000 | 5.135730e+08 | 22662.148687 | 0.1708218 | -0.36455288 | -0.6229618 |
| listings | 1738.02083 | 1581.000 | 1618.5000 | 743.000 | 3296.000 | 2553.000 | 1029.5000 | 5.665690e+05 | 752.707756 | 0.4330833 | 0.64949823 | -0.7917900 |
| months_inventory | 9.19250 | 8.100 | 8.9500 | 3.400 | 14.900 | 11.500 | 3.1500 | 5.306889e+00 | 2.303669 | 0.2506031 | 0.04097527 | -0.1744475 |

Per le altre variabili ho calcolato i principali indici di posizione, di variabilità e di forma. In seguito li ho raggruppati in una tabella di sintesi.

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

```
> rownames(Tabcompl[which.max(Tabcompl$CV),])
[1] "volume"
```

La variabile con variabilità più elevata è “volume” in quanto ha il coefficiente di variazione (CV) più elevato.

```
> rownames(Tabcompl[which.max(abs(Tabcompl$Skewness)),])
[1] "volume
```

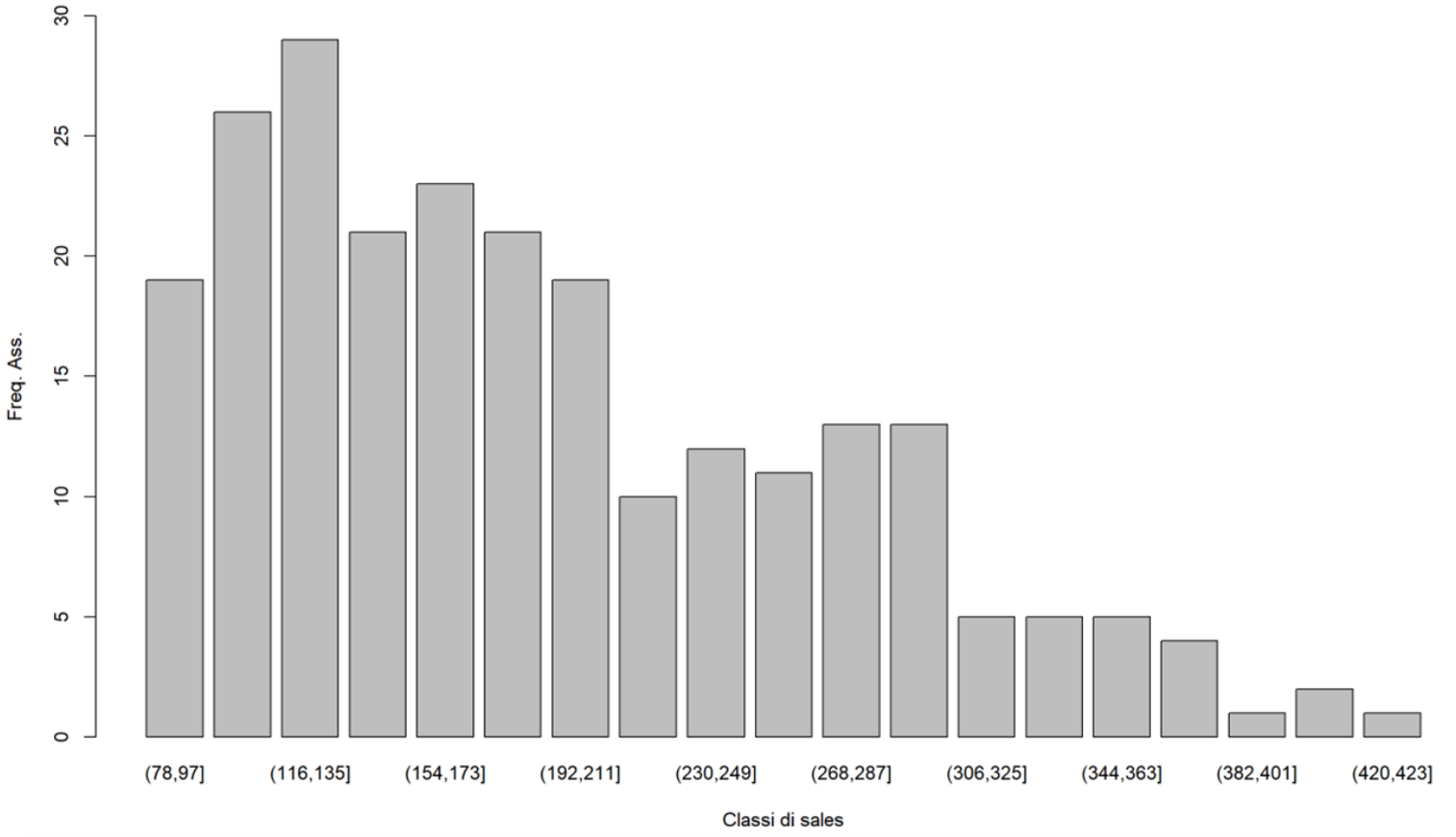
La variabile più asimmetrica è “volume” poiché, in valore assoluto, ha l’indice di asimmetria maggiore.

### 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.

Ho deciso di dividere in classi la variabile “sales”. Ho utilizzato la formula di Sturges per calcolare il numero di classi ed in seguito ho calcolato l’ampiezza di ciascuna classe.

| | ni.class.sales | Ni.class.sales | fi.class.sales | Fi.class.sales |
|---|---|---|---|---|
| (78,97] | 19 | 19 | 0.079166667 | 0.07916667 |
| (97,116] | 26 | 45 | 0.108333333 | 0.18750000 |
| (116,135] | 29 | 74 | 0.120833333 | 0.30833333 |
| (135,154] | 21 | 95 | 0.087500000 | 0.39583333 |
| (154,173] | 23 | 118 | 0.095833333 | 0.49166667 |
| (173,192] | 21 | 139 | 0.087500000 | 0.57916667 |
| (192,211] | 19 | 158 | 0.079166667 | 0.65833333 |
| (211,230] | 10 | 168 | 0.041666667 | 0.70000000 |
| (230,249] | 12 | 180 | 0.050000000 | 0.75000000 |
| (249,268] | 11 | 191 | 0.045833333 | 0.79583333 |
| (268,287] | 13 | 204 | 0.054166667 | 0.85000000 |
| (287,306] | 13 | 217 | 0.054166667 | 0.90416667 |
| (306,325] | 5 | 222 | 0.020833333 | 0.92500000 |
| (325,344] | 5 | 227 | 0.020833333 | 0.94583333 |
| (344,363] | 5 | 232 | 0.020833333 | 0.96666667 |
| (363,382] | 4 | 236 | 0.016666667 | 0.98333333 |
| (382,401] | 1 | 237 | 0.004166667 | 0.98750000 |
| (401,420] | 2 | 239 | 0.008333333 | 0.99583333 |
| (420,423] | 1 | 240 | 0.004166667 | 1.00000000 |

Ci sono 19 classi, delle quali 18 con ampiezza 19 e l’ultima di ampiezza 3.




Ho calcolato l'indice di Gini con il pacchetto "ineq".

```
> ineq(class.sales,type = "Gini")
[1] 0.3627085
```

Il valore dell'indice mostra che c'è solo una lieve disuguaglianza nella distribuzione della variabile "sales".

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

L'indice di Gini per la variabile "city" dovrebbe essere uguale a 1 in quanto tutte le modalità hanno identica frequenza assoluta. Il calcolo effettuato successivamente conferma l'ipotesi iniziale.

### 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?

Per calcolare la probabilità che una riga a caso riporti la città "Beaumont" ho considerato il numero di osservazioni con city=="Beaumont" dividendo per il numero di righe. P=0.25

Per calcolare la probabilità che una riga a caso riporti il mese di Luglio e quindi "7" ho considerato il numero di osservazioni con month==7 dividendo per il numero di righe. P=0.083

Per calcolare la probabilità che una riga a caso riporti l'anno "2012" e il mese di Dicembre "12" ho considerato il numero di osservazioni con years==2012 & month==12 dividendo per il numero di righe. P=0.0167

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

```
> texas$mean_price<-(volume/sales)*100000
```

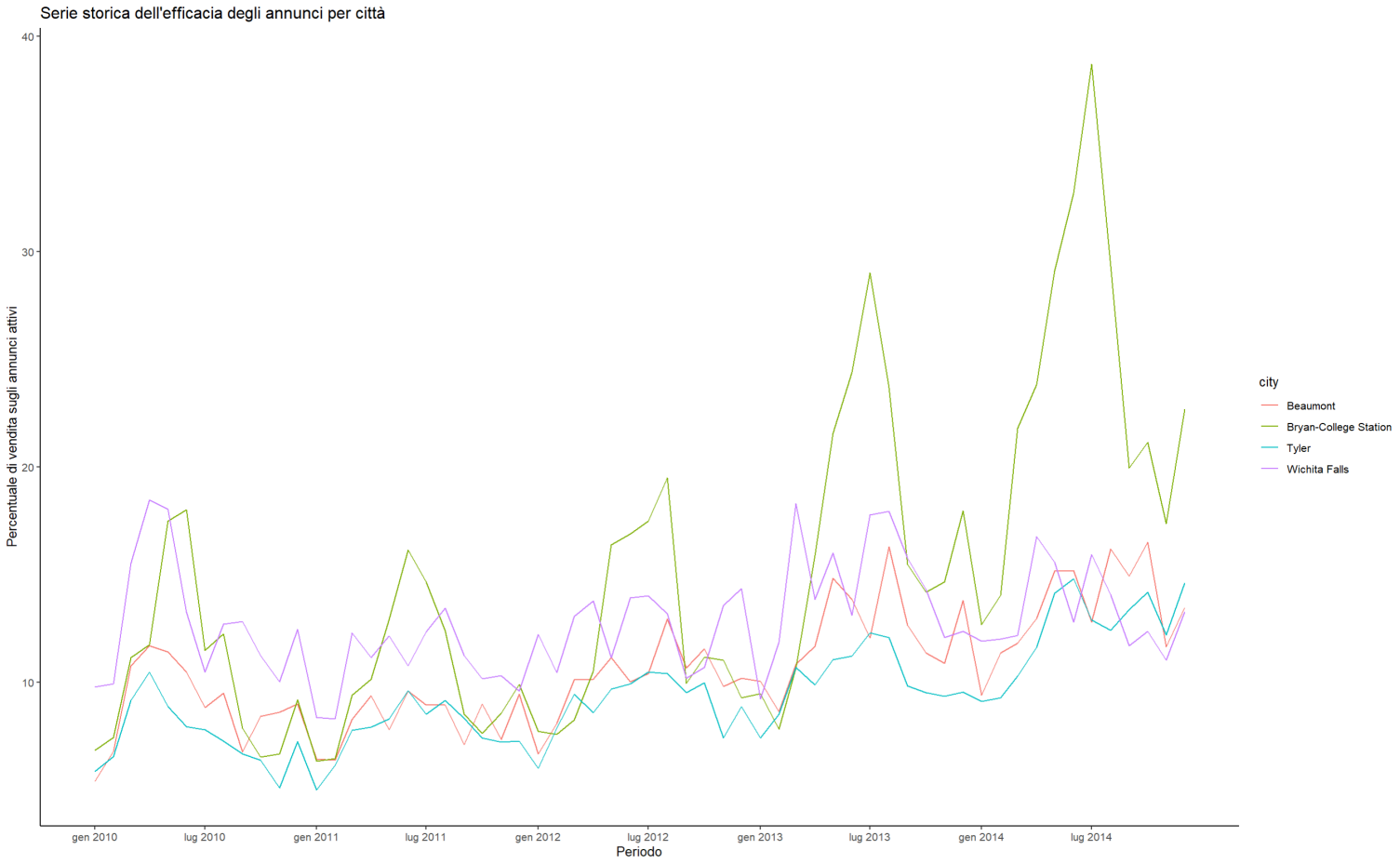
Per creare una colonna che indica il prezzo medio ho calcolato il rapporto tra il valore totale delle vendite e il numero totale di vendite per ciascuna osservazione.

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

```
> texas$efflistings<-(sales/listings)*100
```

Per creare una colonna che rifletta l'efficacia degli annunci di vendita ho calcolato, per ogni osservazione, la percentuale di annunci che hanno portato ad un'effettiva vendita.

In seguito ho realizzato un grafico che mostra l'efficacia delle vendite in ogni città per ogni mese dal 2010 al 2014:




Ci sono due considerazioni da fare:

1) Gli annunci di vendita più efficaci sono quelli presenti nella città Bryan-College Station;

2) In generale si osserva che i trend positivi e negativi si verificano contemporaneamente nelle quattro città ma con magnitudine diversa.

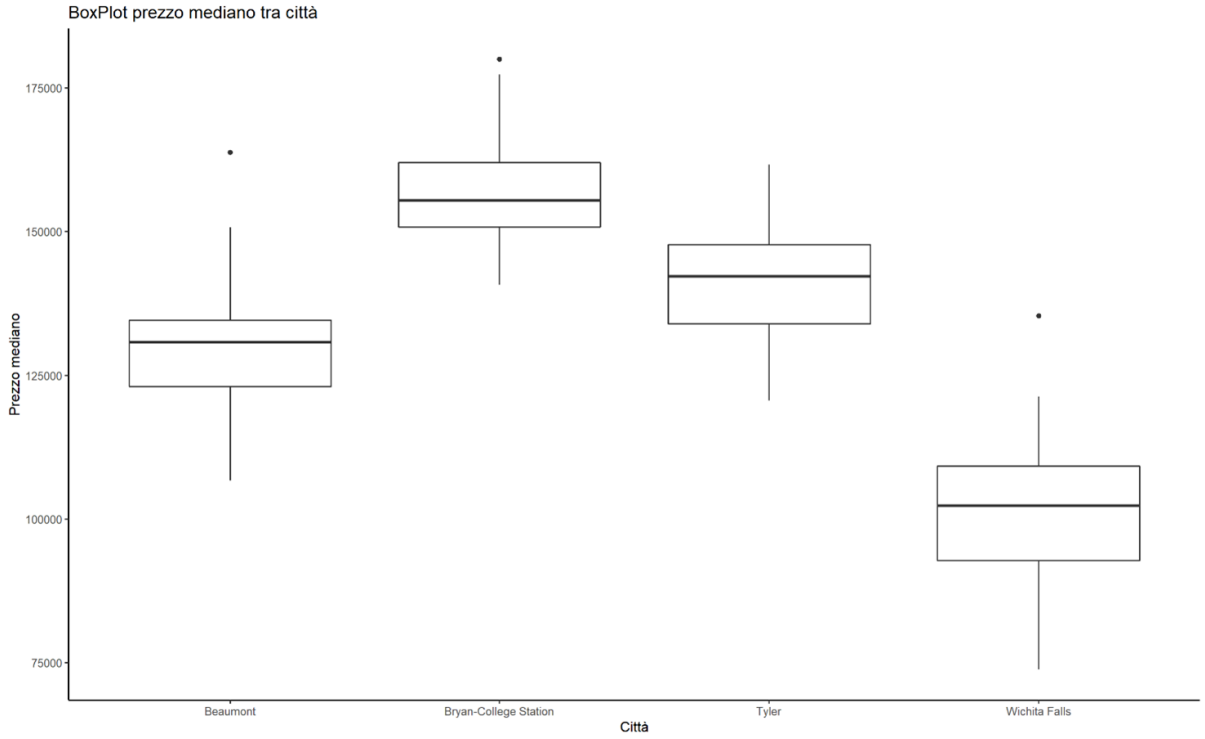
**10) Prova a creare dei summary(), o semplicemente media e deviazione standard, di alcune variabili a tua scelta, condizionatamente alla città, agli anni e ai mesi. Puoi utilizzare il linguaggio R di base oppure essere un vero Pro con il pacchetto dplyr.**

```
> texas %>%
+   group_by(year,city) %>%
+   summarise(media.di.sales=mean(sales),
+             sd.di.volume=sd(volume))
```

| year | city | media.di.sales | sd.di.volume |
|---|---|---|---|
| <int> | <chr> | <dbl> | <dbl> |
| 1 | 2010 | Beaumont | 156. | 4.95 |
| 2 | 2010 | Bryan-College Station | 168. | 10.8 |
| 3 | 2010 | Tyler | 228. | 8.39 |
| 4 | 2010 | Wichita Falls | 123. | 4.07 |
| 5 | 2011 | Beaumont | 144 | 4.30 |
| 6 | 2011 | Bryan-College Station | 167. | 10.3 |
| 7 | 2011 | Tyler | 239. | 9.41 |
| 8 | 2011 | Wichita Falls | 106. | 2.52 |
| 9 | 2012 | Beaumont | 172. | 4.92 |
| 10 | 2012 | Bryan-College Station | 197. | 13.5 |
| 11 | 2012 | Tyler | 264. | 10.2 |
| 12 | 2012 | Wichita Falls | 112. | 2.66 |
| 13 | 2013 | Beaumont | 201. | 6.44 |
| 14 | 2013 | Bryan-College Station | 238. | 19.5 |
| 15 | 2013 | Tyler | 287. | 10.3 |
| 16 | 2013 | Wichita Falls | 121. | 3.11 |
| 17 | 2014 | Beaumont | 214. | 7.05 |
| 18 | 2014 | Bryan-College Station | 260. | 18.0 |
| 19 | 2014 | Tyler | 332. | 12.8 |
| 20 | 2014 | Wichita Falls | 117 | 3.13 |

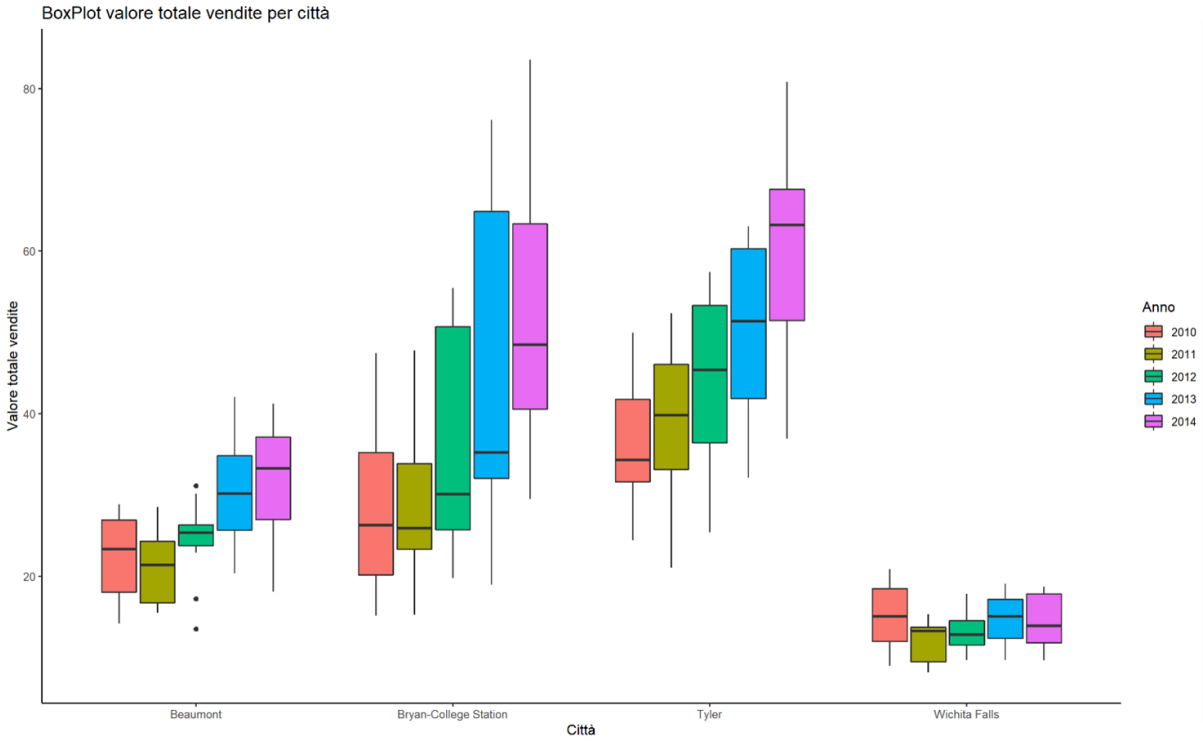
**GRAFICI GGPLOT2**

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



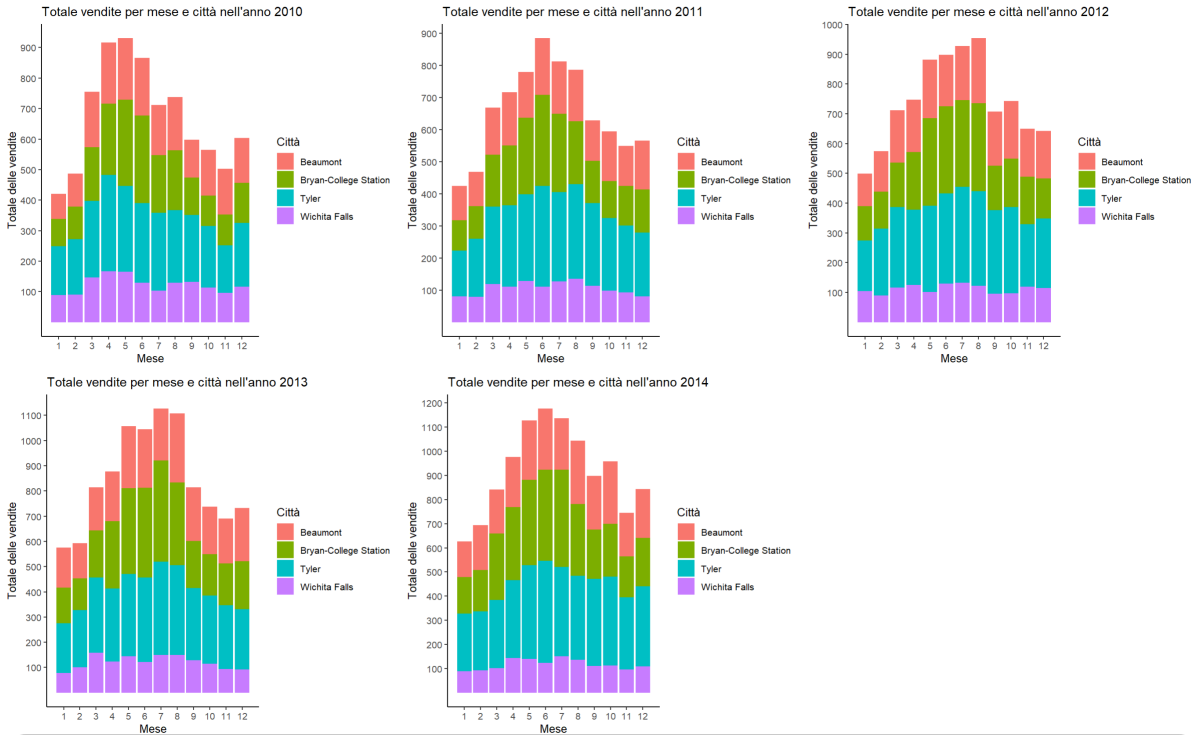

Dal grafico si evince che la variabile “median_price” presenta una maggiore variabilità per la città Wichita Falls, dove ci sono anche case con i prezzi minori, mentre la stessa variabile si distribuisce con una minore variabilità per le osservazioni relative alla città Bryan-College Station, dove sono presenti le case più costose.

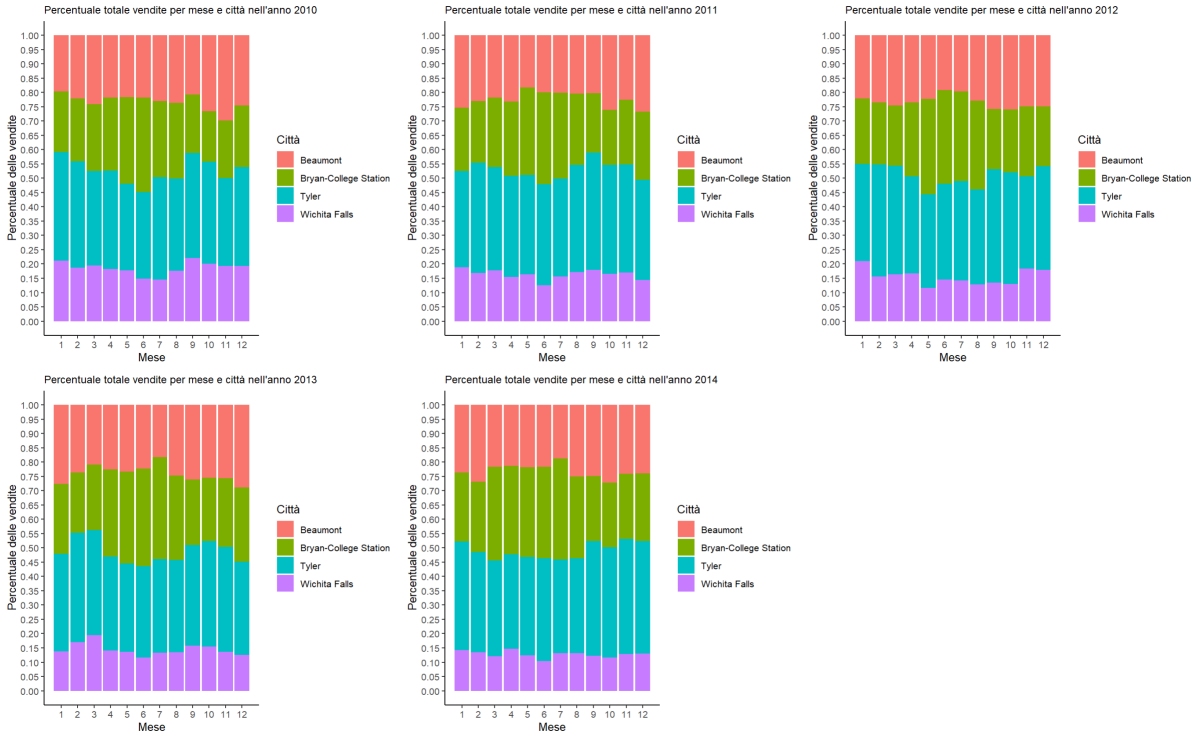
**2) Utilizza i boxplot o qualche variante per confrontare la distribuzione del valore totale delle vendite tra le varie città ma anche tra i vari anni. Qualche considerazione da fare?**




Dal grafico si può osservare che la variabile “volume” è caratterizzata da variabilità maggiore quando si tratta della città Bryan-College Station. In particolare il valore totale delle vendite ha una variabilità maggiore nell’anno 2013. Invece “volume” è meno variabile per la città Wichita Falls. Infine nell’anno 2012, limitatamente alla città Beumont, il volume delle vendite si caratterizza, in assoluto, per una minore variabilità.

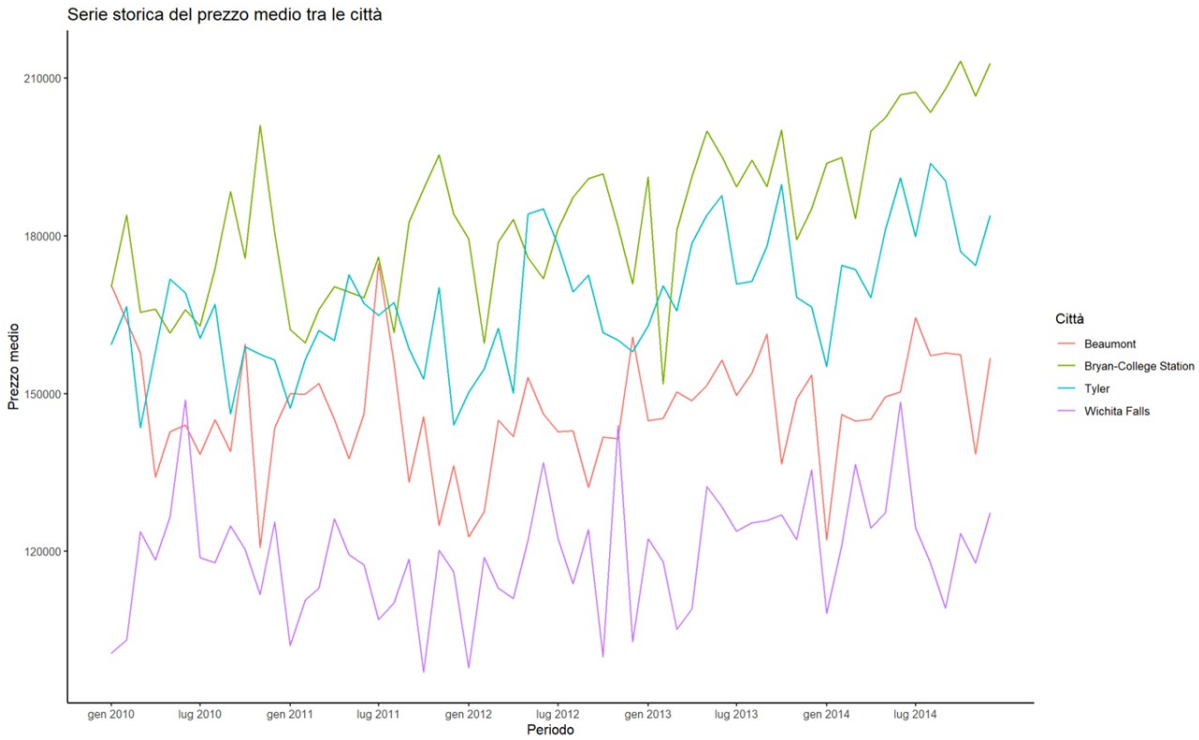
**3) Usa un grafico a barre sovrapposte per ogni anno, per confrontare il totale delle vendite nei vari mesi, sempre considerando le città. Prova a commentare ciò che viene fuori. Già che ci sei prova anche il grafico a barre normalizzato. Consiglio: Stai attento alla differenza tra geom_bar() e geom_col(). PRO LEVEL: cerca un modo intelligente per inserire ANCHE la variabile Year allo stesso blocco di codice, senza però creare accrocchi nel grafico.**






Dai grafici si nota che per gran parte dei mesi presi in esame la percentuale di vendite più alta si registra nella città di Tyler mentre la più bassa nella città di Wichita Falls. Inoltre le percentuali di vendita per ogni città risultano essere più o meno costanti nel tempo.

**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**




Ho scelto di rappresentare il prezzo medio delle vendite in ciascun mese dal 2010 al 2015 per le città presenti nel dataset. Si vede che i prezzi medi più alti si verificano nella città di Bryan-College Station mentre i più bassi riguardano la città di Wichita Falls.