Analisi vendite Texas

Questo documento ha lo scopo di documentare l'analisi dei del dataset relativo alle vendite di case in Texas.

Esplorazione

Iniziamo osservando una piccola parte del dataset preso in analisi.

	city	year	month	sales	volume	median_price	listings	months_inventory
1	Beaumont	2010	1	83	14.162	163800	1533	9.5
2	Beaumont	2010	2	108	17.690	138200	1586	10.0
3	Beaumont	2010	3	182	28.701	122400	1689	10.6
4	Beaumont	2010	4	200	26.819	123200	1708	10.6
5	Beaumont	2010	5	202	28.833	123100	1771	10.9
6	Beaumont	2010	6	189	27.219	122800	1803	11.1
7	Beaumont	2010	7	164	22.706	124300	1857	11.7
8	Beaumont	2010	8	174	25.237	136800	1830	11.6
9	Beaumont	2010	9	124	17.233	121100	1829	11.7
10	Beaumont	2010	10	150	23.904	138500	1779	11.5

Abbiamo le seguenti variabili:

- city: città **QA**
- year: anno di riferimento **QU**
- month: mese di riferimento **QU**
- sales: numero totale di vendite **QU**
- volume: valore totale delle vendite in milioni di dollari **QU**
- median_price: prezzo mediano di vendita in dollari **QU**
- listings: numero totale di annunci attivi **QU**
- months_inventory: quantità di tempo necessaria per vendere tutte le inserzioni correnti al ritmo attuale delle vendite, espresso in mesi **QU**

Che hanno i seguenti tipi
QA: Variabile Qualitativa
QU: Variabile Quantitativa

Analisi delle variabili

Analizziamo ora le singole variabili una a una evidenziano i vari indici.

City

Osservando le **distribuzione di frequenza** capiamo subito che la **distribuzione delle classi è eterogenea** per le quattro città presenti. Per confermare questa ipotesi calcoliamo **l'indice di eterogeneità di Gini normalizzato che è pari a 1** ovvero il massimo grado di eterogeneità possibile.

	ni	fi
Beaumont	60	0.25
Bryan-College Station	60	0.25
Tyler	60	0.25
Wichita Falls	60	0.25

Questa tabella corrisponde anche alla distribuzione di probabilità della classe city, quindi per rispondere alla domanda *"Qual è la probabilità che presa una riga a caso di questo dataset essa riporti la città "Beaumont ?*" Ci basta guardare la tabella. La probabilità sarà quindi 0.25

Per completezza costruiamo diverse distribuzioni di frequenza doppie. Relative ad anno e mese di vendita. Anche in questi casi Osserviamo che non ci sono informazioni particolarmente interessanti e le vendite sono sparse in modo omogeneo per i mesi e anni.

	year				
city	2010	2011	2012	2013	2014
Beaumont	12	12	12	12	12
Bryan-College Station	12	12	12	12	12
Tyler	12	12	12	12	12
Wichita Falls	12	12	12	12	12

	month											
city	1	2	3	4	5	6	7	8	9	10	11	12
Beaumont	5	5	5	5	5	5	5	5	5	5	5	5
Bryan-College Station	5	5	5	5	5	5	5	5	5	5	5	5
Tyler	5	5	5	5	5	5	5	5	5	5	5	5
Wichita Falls	5	5	5	5	5	5	5	5	5	5	5	5

Year

Osserviamo ora la variabile anni e notiamo che le registrazioni partono dall'anno 2010 e arrivano fino all'anno 2014. Anche in questo caso i dati sono distribuiti eterogenea mente per tutti gli anni delle registrazioni, avendo un 20% di vendite all'anno.

	ni_year	fi_year
2010	48	0.2
2011	48	0.2
2012	48	0.2
2013	48	0.2
2014	48	0.2

Calcoliamo anche in questo caso **l'indice di Gini normalizzato, che sarà pari a 1.**

Month

La variabile mesi è molto simile alla variabile anni, costruendo la distribuzione di frequenza possiamo notare che le vendite oltre a essere distribuite in modo eterogeneo negli anni, lo sono anche nei mesi.
Abbiamo quindi 4 vendite al mese, per un totale di 48 vendite all'anno.
L'indice di Gini normalizzato è pari a 1

	ni_month	fi_month
1	20	0.08333333
2	20	0.08333333
3	20	0.08333333
4	20	0.08333333
5	20	0.08333333
6	20	0.08333333
7	20	0.08333333
8	20	0.08333333
9	20	0.08333333
10	20	0.08333333
11	20	0.08333333
12	20	0.08333333

Questa tabella corrisponde anche alla distribuzione di probabilità e quindi se volessimo rispondere alla domanda *"Quale la probabilità scegliendo a caso del dataset che il mese corrisponda a luglio ?"→ la probabilità è di 0.08.*

Distribuzione Congiunta Anno-Mese

year	month 1	2	3	4	5	6
2010	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667
2011	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667
2012	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667
2013	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667
2014	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667

year	month 7	8	9	10	11	12
2010	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667
2011	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667
2012	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667
2013	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667
2014	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667	0.01666667

Costruiamo ora una tabella di frequenze doppia Anni-Mesi. La tabella corrisponderà alla distribuzione di probabilità dei mesi negli anni, possiamo quindi rispondere alla domanda "*Quale la probabilità scegliendo a caso del dataset che il mese corrisponda a dicembre dell'anno 2012 ?*" → La risposta sarà esattamente 0.016.

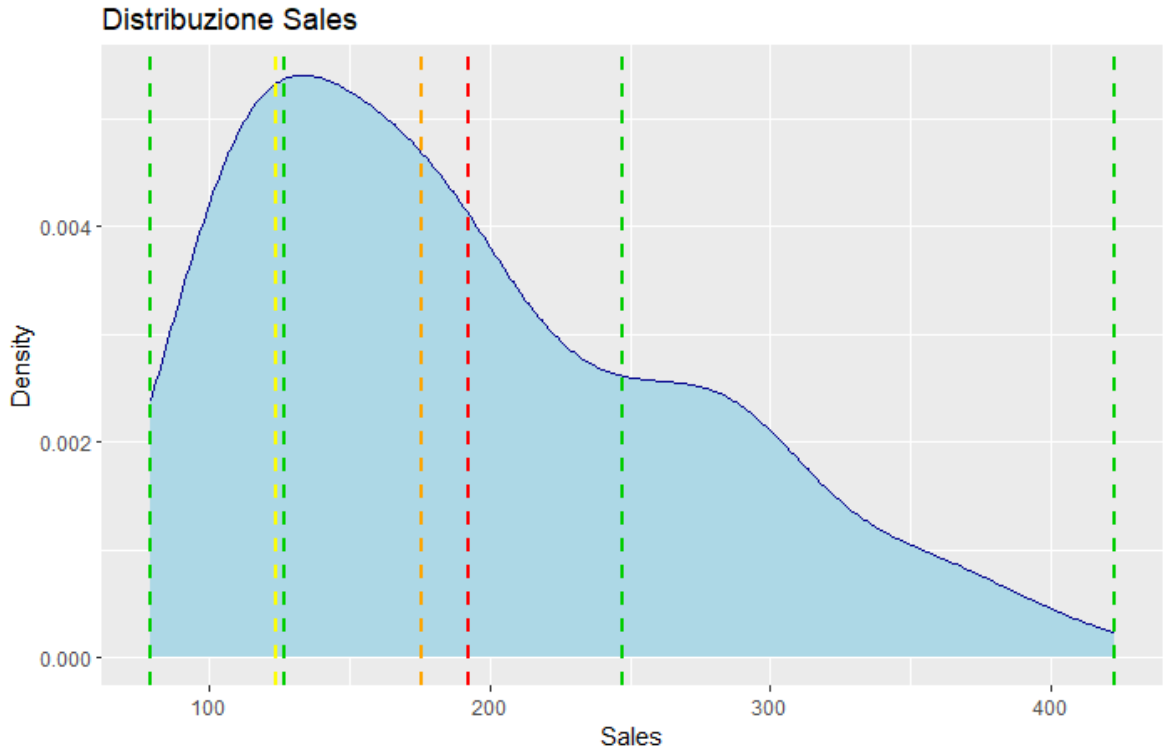
Sales

La variabile vendite è una delle più interessanti della nostra analisi.
Iniziando osservando i vari indici di posizione

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
79.0	127.0	175.5	192.3	247.0	423.0

Possiamo notare che i valori vanno da un **minimo di 79** a **massimo di 423** con un **range pari a 344** e un **range interquantile pari a 120.**
Inoltre la **moda è pari a 124.**
Dalle informazioni degli indici di posizione possiamo ipotizzare dato che la media > mediana > moda ci troviamo davanti a una distribuzione asimmetrica positiva. Osserviamo il grafico della distribuzione per avere più informazioni.

In Rosso abbiamo la media, in giallo la moda, in arancione la mediana/secondo quartile e in verde i cinque quartili.

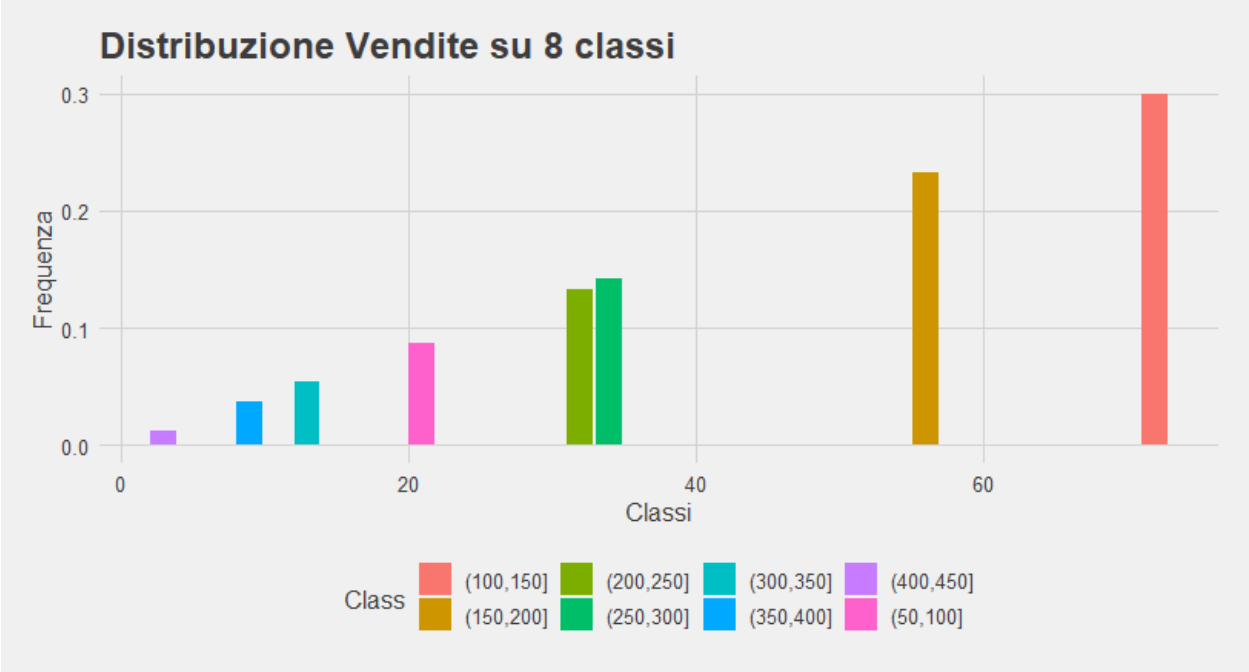


Per assicurarci di questa assimetria positiva calcoliamo indici di variabili e forma.
Abbiamo una **varianza pari a 6344.4**, una **deviazione standard pari a 79.65** e un **coefficiente di variazione pari a 41.42.** Grazie a questi valori calcoliamo il **Coefficiente di simmetria che è pari a 0.72** e la **Curtosi che è pari a -0.31.** Questi due valori ci conferano che la distribuzione è **assimetrica positiva**, ma ci dicono anche che la **distribuzione è platicurtica**, ovvero più appiattita rispetto a una distribuzione normale.

Raggrumiamo ora le vendite in classi, questo ci permetterà di osservare i fatti da un altro punto di vista. *La prima divisione che facciamo consiste nel dividere le vendite in otto classi, ognuna con un range di 50; partiremo dal valore 50 e arriveremo al valore 450.* Costruendo la distribuzione di frequenza di questa classe otteniamo.

	ni	fi	Ni	Fi
(50,100]	21	0.08750000	21	0.0875000
(100,150]	72	0.30000000	93	0.3875000
(150,200]	56	0.23333333	149	0.6208333
(200,250]	32	0.13333333	181	0.7541667
(250,300]	34	0.14166667	215	0.8958333
(300,350]	13	0.05416667	228	0.9500000
(350,400]	9	0.03750000	237	0.9875000
(400,450]	3	0.01250000	240	1.0000000

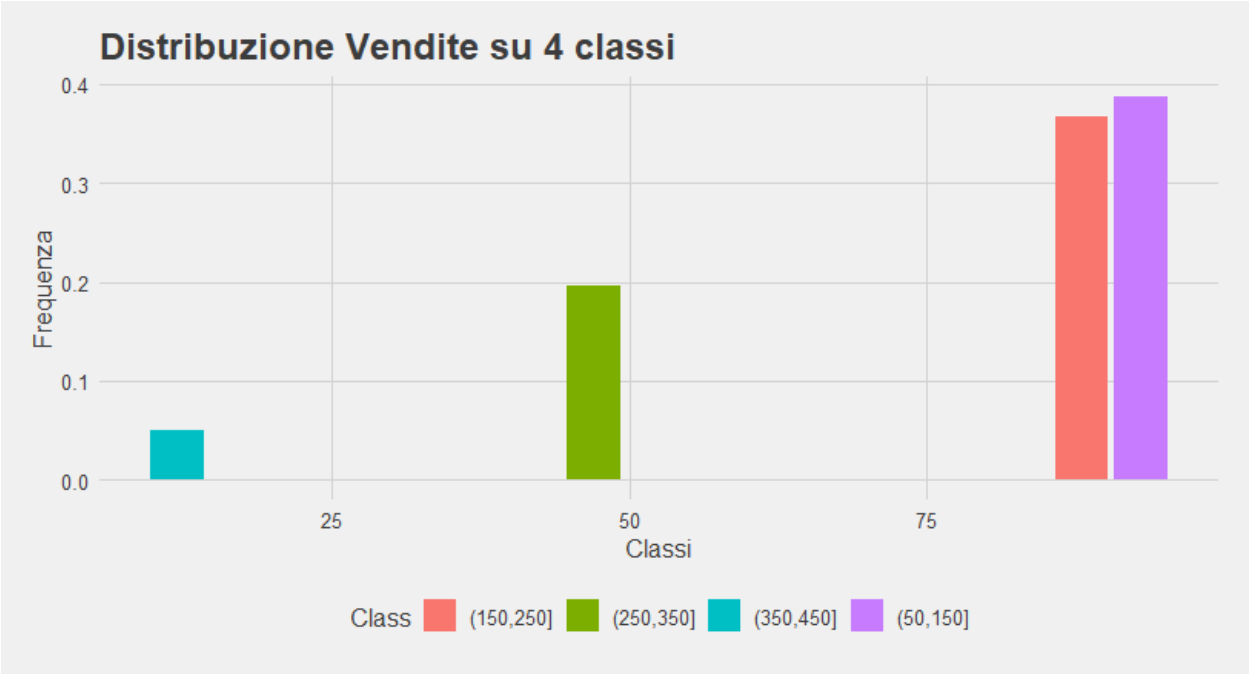
Possiamo notare che la **moda è la classe (100,150]**. Per questa divisione in classi abbiamo un **indice di Gini normalizzato pari a 0.9206.** Osserviamo ora questa distribuzione in classi con un grafico a barre.



Ma facciamo un altro test, *dividiamo le vendite in sole quattro classi che copriranno un range di 100 valori, sempre partendo da 50 e arrivando a 450.* Costruita la distribuzione di frequenza abbiamo che:

	ni	fi	Ni	Fi
(50,150]	93	0.3875000	93	0.3875000
(150,250]	88	0.3666667	181	0.7541667
(250,350]	47	0.1958333	228	0.9500000
(350,450]	12	0.0500000	240	1.0000000

La **moda sarà la classe che va da (50,150]**. **L'indice di Gini normalizzato per questa divisione in classi è pari a 0.8996.** Osserviamo ora il grafico a barre di questa distribuzione in classi.

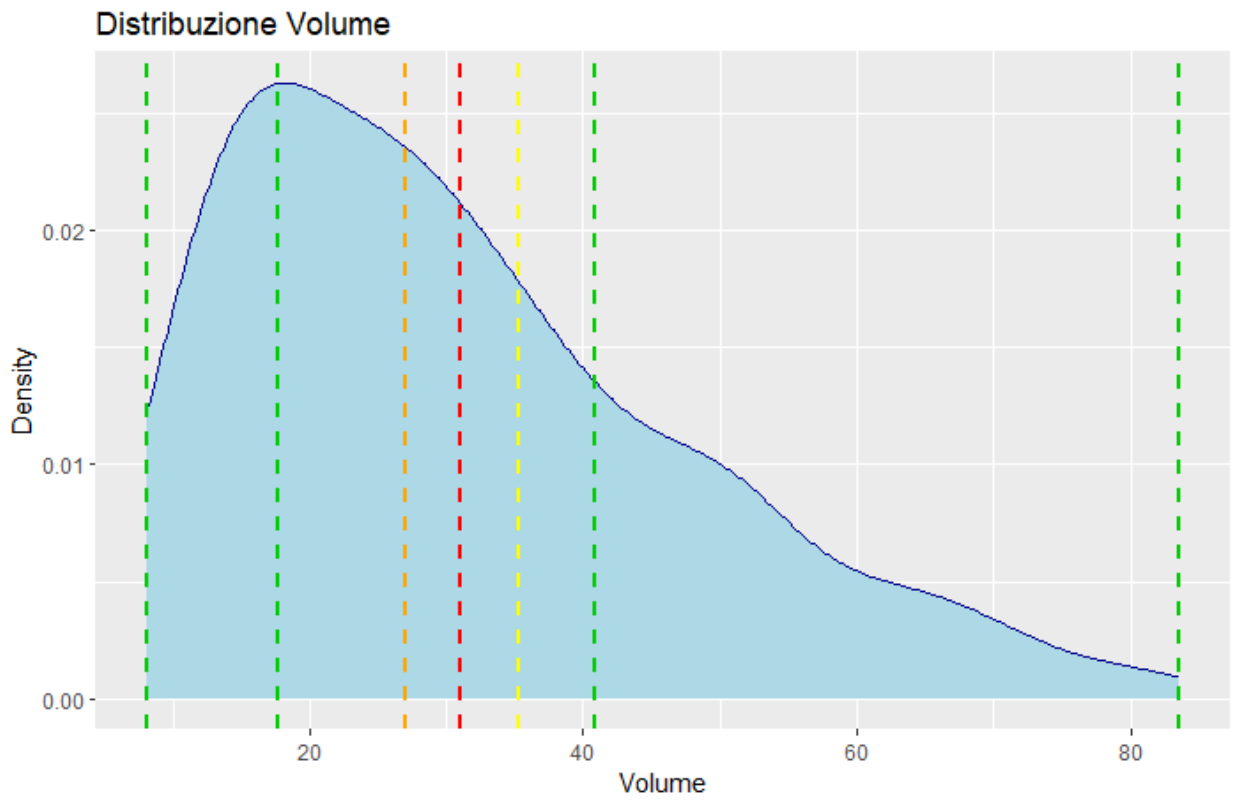


Volume

Iniziamo osservando gli indici di posizione.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
8.166	17.660	27.062	31.005	40.893	83.547

Abbiamo un valore **minimo pari a 8.166** e un **massimo pari a 83.547**. Un **range pari a 73.38** e un **range interquantile pari a 23.23,** inoltre la **moda è pari a 35.33.** Già da queste informazioni possiamo ipotizzare che la distribuzione assimetrica positiva e a breve lo verificheremo.
Osserviamo i grafico della distribuzione:
In Rosso abbiamo la media, in giallo la moda, in arancione la mediana/secondo quartile e in verde i cinque quartili.



Anche dalla forma della distribuzione sembrerebbe una distribuzione asimmetrica positiva. Per Confermarlo calcoliamo indici di variabilità e forma.
Avremo una **varianza pari a 277.27**, una **deviazione standard pari a 16.65** e un **coefficiente di variazione pari a 53.70**.
Infine calcoliamo **indice di simmetria che è pari a 0.88** e la **curtosi che è pari a 0.18.** Grazie a questi indici possiamo confermare che siamo di fronte, a una **distribuzione asimmetrica positiva leptocurtica.**

Median price

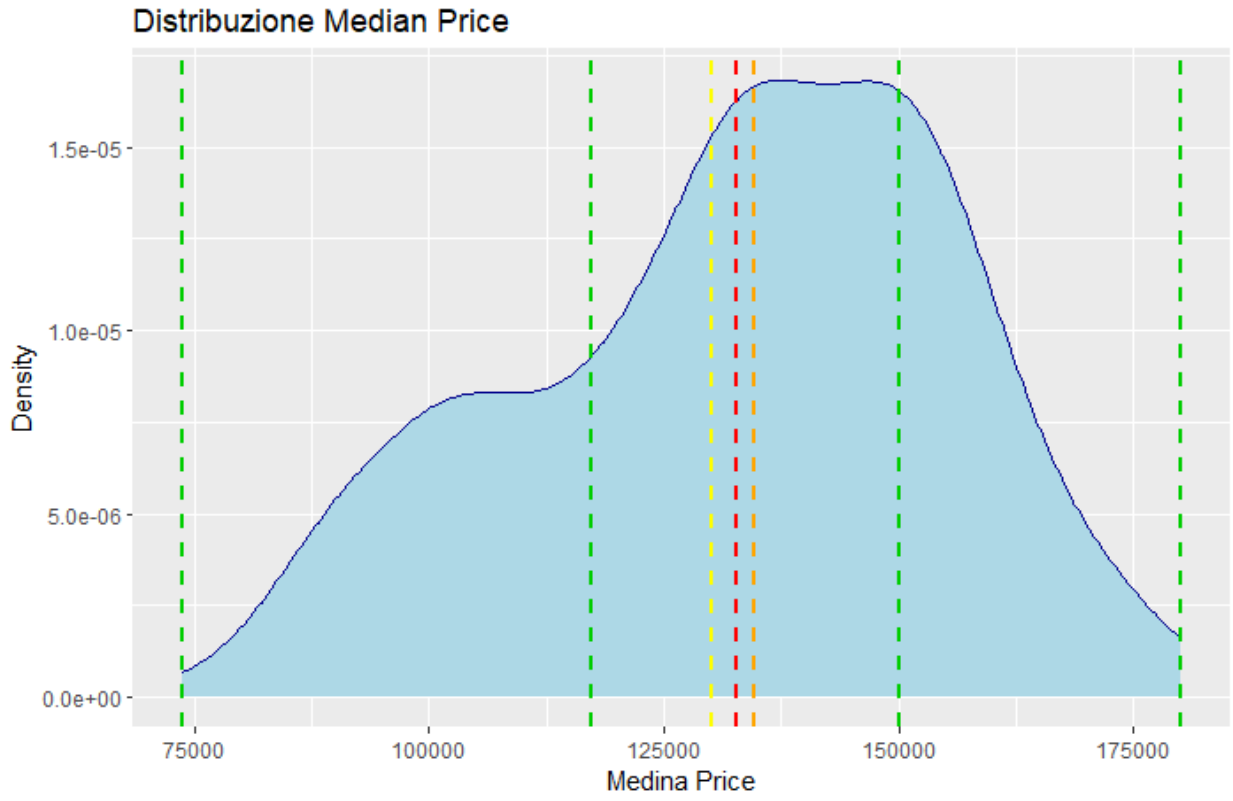
Iniziamo osservando gli indici di posizione.

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 73800  117300  134500  132665  150050  180000
```

Abbiamo un valore **minimo pari a 73800** e un **massimo pari a 180000**. Un **range pari a 106200** e un **range interquantile pari a 32750,** inoltre la **moda è pari a 130000**. Dato che la moda>mediana>media sappiamo che si tratta di una distribuzione asimmetrica negativa.

Osserviamo il grafico della distribuzione:

In Rosso abbiamo la media, in giallo la moda, in arancione la mediana/secondo quartile e in verde i cinque quartili.



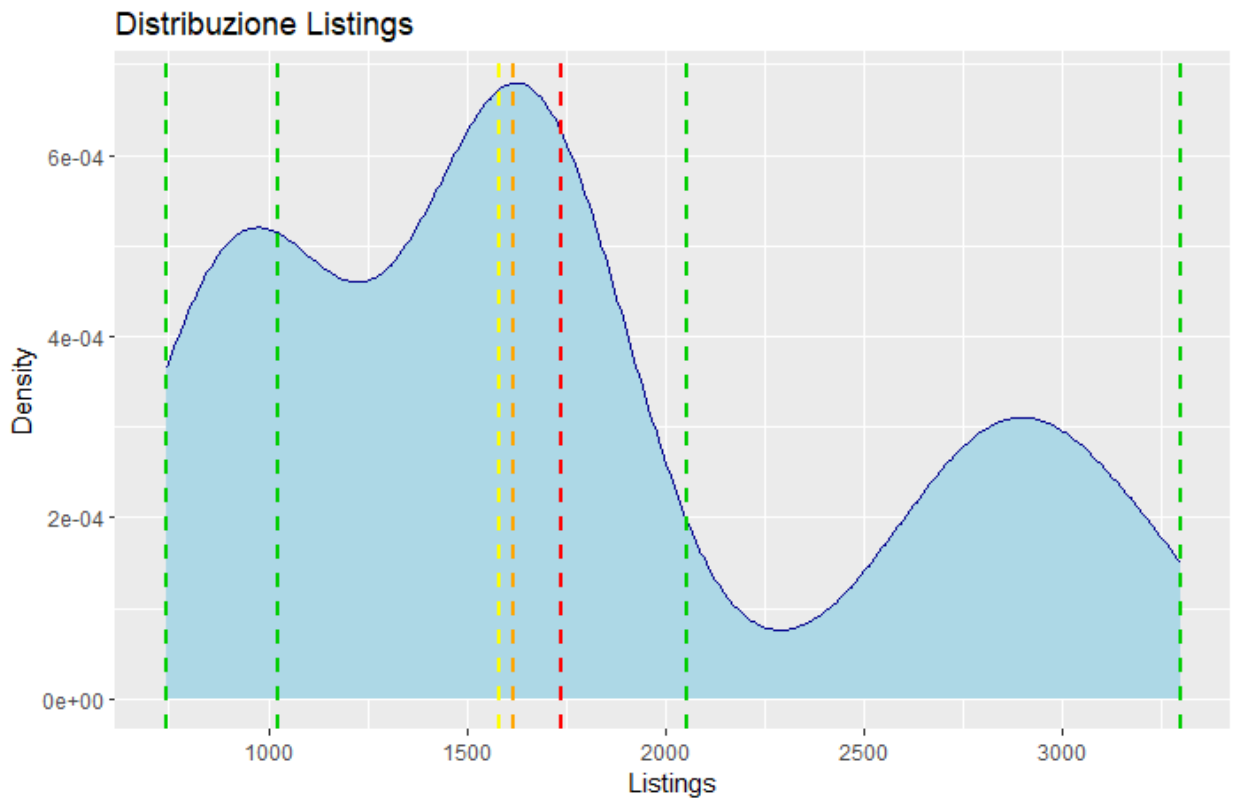
Calcoliamo ora i vari indici di variabilità e forma.
Avremo una **varianza pari a 5135729883**, una **deviazione standard pari a 22662** e un **coefficiente di variazione pari a 17.08**.
Infine calcoliamo **indice di simmetria che è pari a -0.3645** e la **curtosi che è pari a -0.6230.**
Confermiamo che si tratta i una distribuzione asimmetrica negativa e che inoltre è platicurtica.

Listing

Iniziamo osservando gli indici di posizione.

```
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 743    1026    1618    1738    2056    3296
```

Abbiamo un valore **minimo pari a 743** e un **massimo pari a 3296**. Un **range pari a 2553** e un **range interquantile pari a 1029.5,** inoltre la **moda è pari a 1581**. Dato che la moda<mediana<media sappiamo che si tratta di una distribuzione asimmetrica positiva.
Osserviamo i grafico della distribuzione:
In Rosso abbiamo la media, in giallo la moda, in arancione la mediana/secondo quartile e in verde i cinque quartili.



Calcoliamo ora i vari indici di variabilità e forma.
Avremo una **varianza pari a 566568.96**, una **deviazione standard pari a 752.70** e un **coefficiente di variazione pari a 43.30**.
Infine calcoliamo **indice di simmetria che è pari a 0.6495** e la **curtosi che è pari a -0.7918.**
Confermiamo che si tratta i una distribuzione asimmetrica positivo e che inoltre è platicurtica.

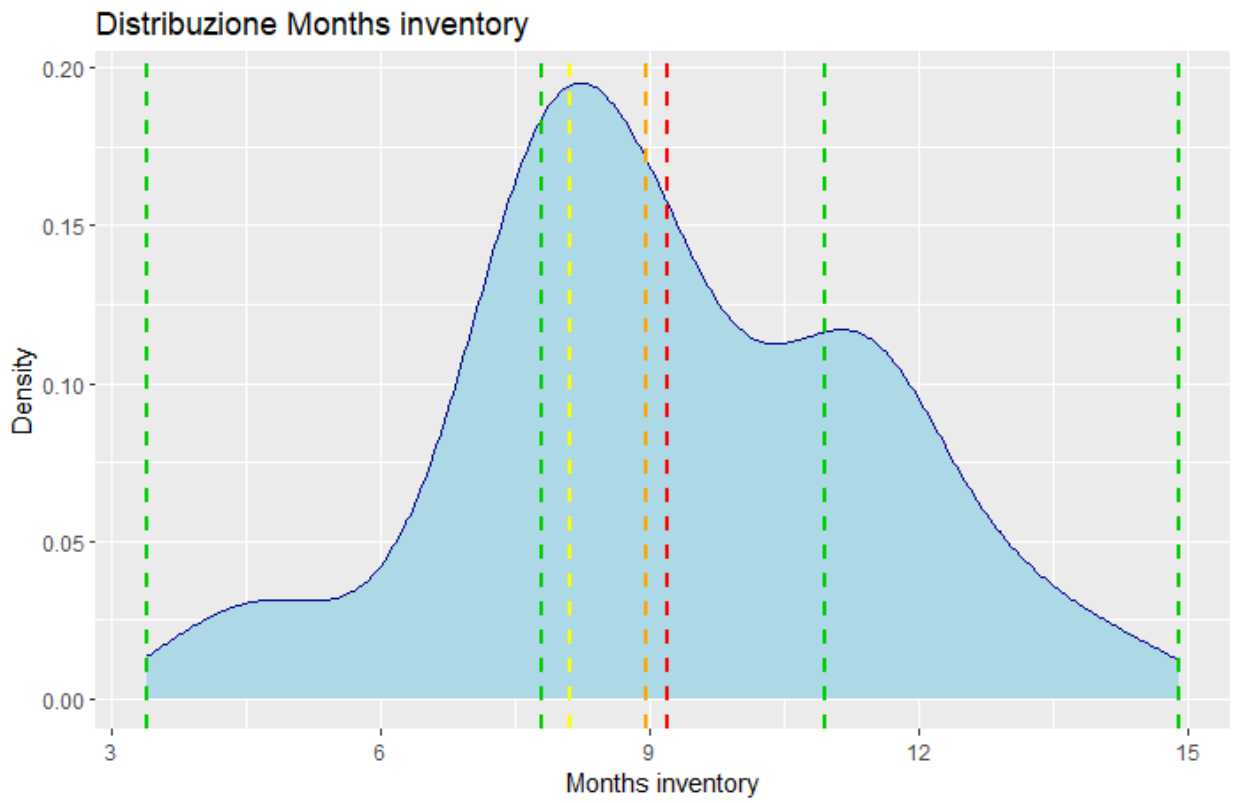
Months inventory

Iniziamo osservando gli indici di posizione.

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
3.400	7.800	8.950	9.193	10.950	14.900

Abbiamo un valore **minimo pari a 3.400** e un **massimo pari a 14.900**. Un **range pari a 11.5** e un **range interquantile pari a 3.15,** inoltre la **moda è pari a 8.1**. Dato che la moda<mediana<media sappiamo che si tratta di una distribuzione asimmetrica positiva.

Osserviamo i grafico della distribuzione:
In Rosso abbiamo la media, in giallo la moda, in arancione la mediana/secondo quartile e in verde i cinque quartili.



Calcoliamo ora i vari indici di variabilità e forma.
Avremo una **varianza pari a 5.3069**, una **deviazione standard pari a 2.3037** e un **coefficiente di variazione pari a 5.3069**.
Infine calcoliamo **indice di simmetria che è pari a 0.0409** e la **curtosi che è pari a -0.1744**
Confermiamo che si tratta i una distribuzione asimmetrica positivo e che inoltre è platicurtica.

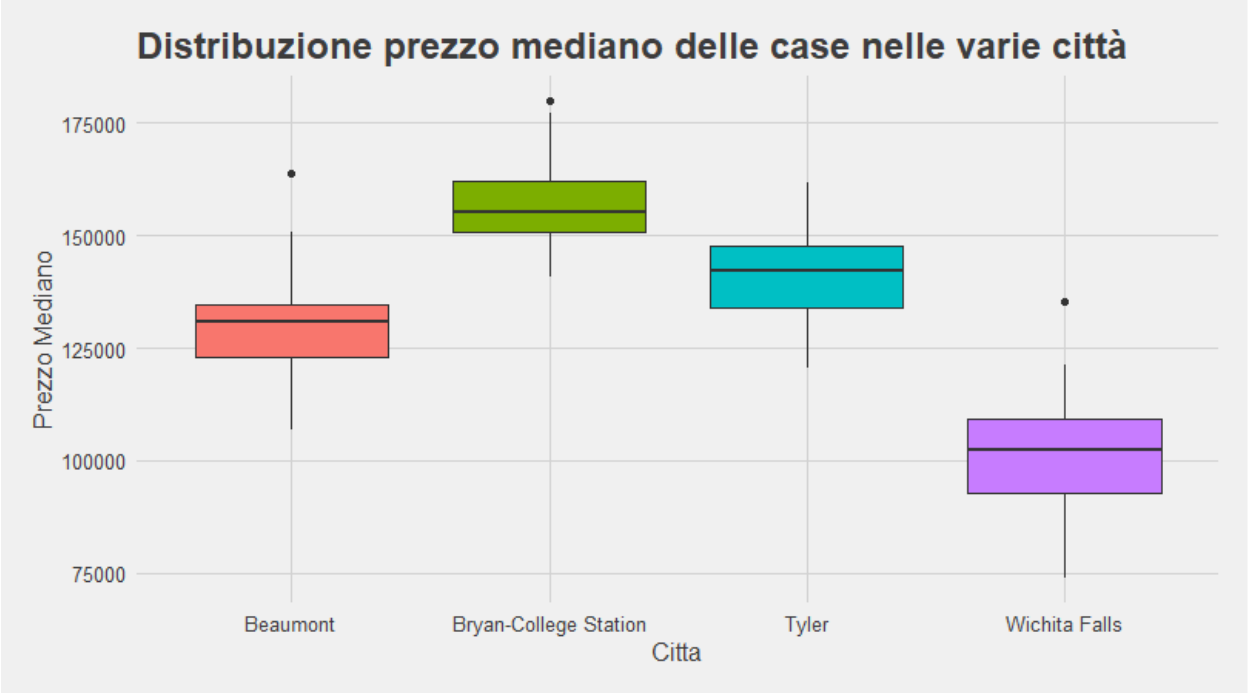
Confronti

Iniziamo ora a confrontare le variabili mostrando una tabella di tutti gli indici di posizione, variabilità e forma.
Possiamo notare che la variabile con variabilità più elevata è le vendite, la variabile più asimmetrica è il volume avendo una coda molto allungata verso destra.

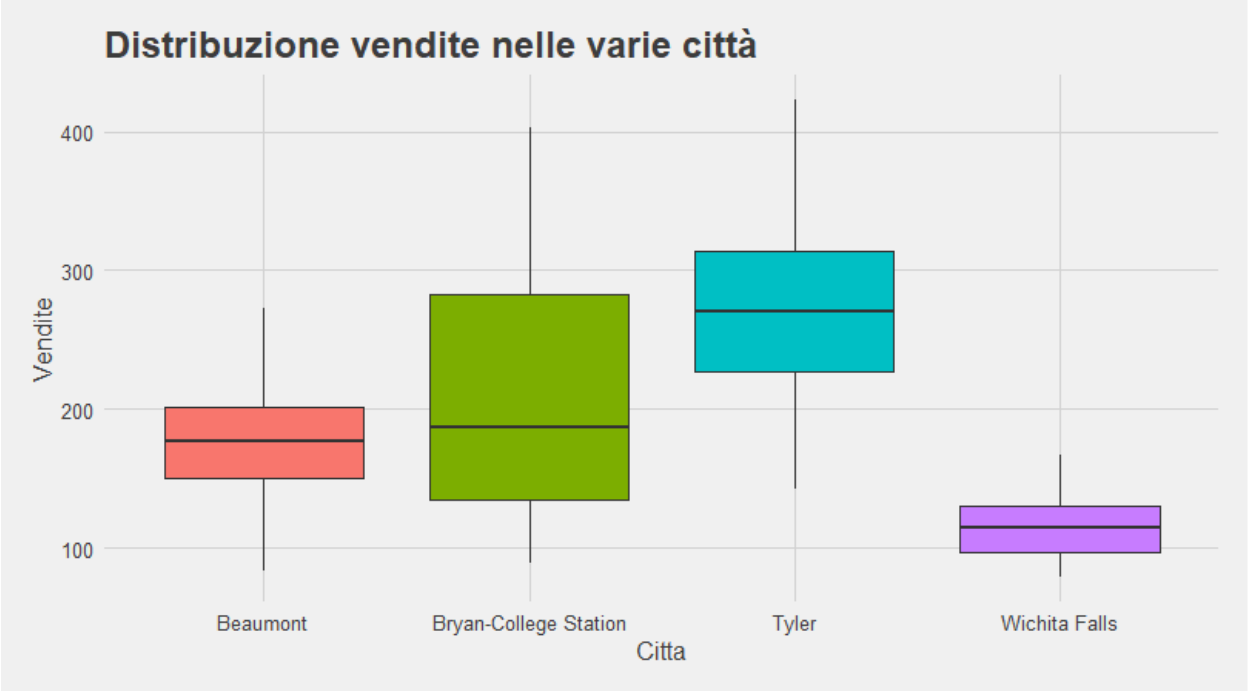
	sales_summary	volume_summary	median_price_summary	listings_summary	months_inventory_summary
Min.	79	8.166	73800	743	3.4
1st Qu.	127	17.6595	117300	1026.5	7.8
Median	175.5	27.0625	134500	1618.5	8.95
Mean	192.291666666667	31.0051875	132665.416666667	1738.02083333333	9.1925
3rd Qu.	247	40.893	150050	2056	10.95
Max.	423	83.547	180000	3296	14.9
Range	344	75.381	106200	2553	11.5
IQR	120	23.2335	32750	1029.5	3.15
Mode	124	35.335	130000	1581	8.1
Var	6344.29951185495	277.270692404027	513572983.089261	566568.966091353	5.30688912133891
SD	79.6511111777793	16.6514471564494	22662.148686505	752.707756098841	2.30366862229334
CV	41.4220296482492	53.7053586805415	17.0821825732064	43.3083275909432	25.0603059264982
Asymmetry	0.718104024884959	0.884742026325995	-0.364552878177372	0.649498226273971	0.040975265871081
Curtosi	-0.313176409071494	0.176986997089741	-0.622961820755544	-0.791790033332591	-0.174447541638487

BoxPlot

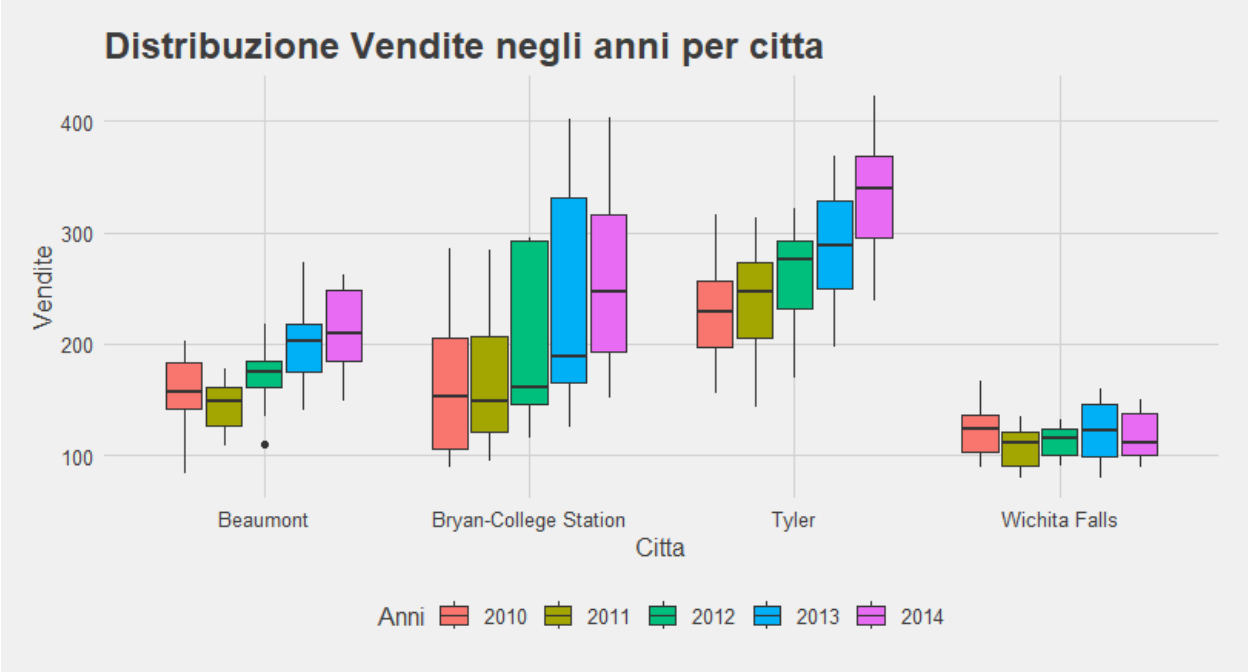
Osserviamo alcuni confronti sulle distribuzioni utilizzando i boxplot.
Iniziamo analizzando le differenze di distribuzione del prezzo mediano per le varie città



Notiamo fin da subito che Bryan-College Station ha i prezzi mediani più alti, e che quindi in generale i prezzi delle case siano più alti, questa ipotesi verrà osservata meglio nella domanda “Quale la città con il prezzo più alto ?” In cui viene vengono calcolati i prezzi medi delle case e mostrati su un grafico.
Osserviamo poi la distribuzione delle vendite nelle varie città.



Questo grafico ha però un problema ovvero che comprime tutti gli anni e mesi e quindi mostra la distribuzione generale. Per rendere il tutto più chiaro Osserviamo una distribuzione divisa per anni, in cui avremo le vendite di ogni città per ogni anno.

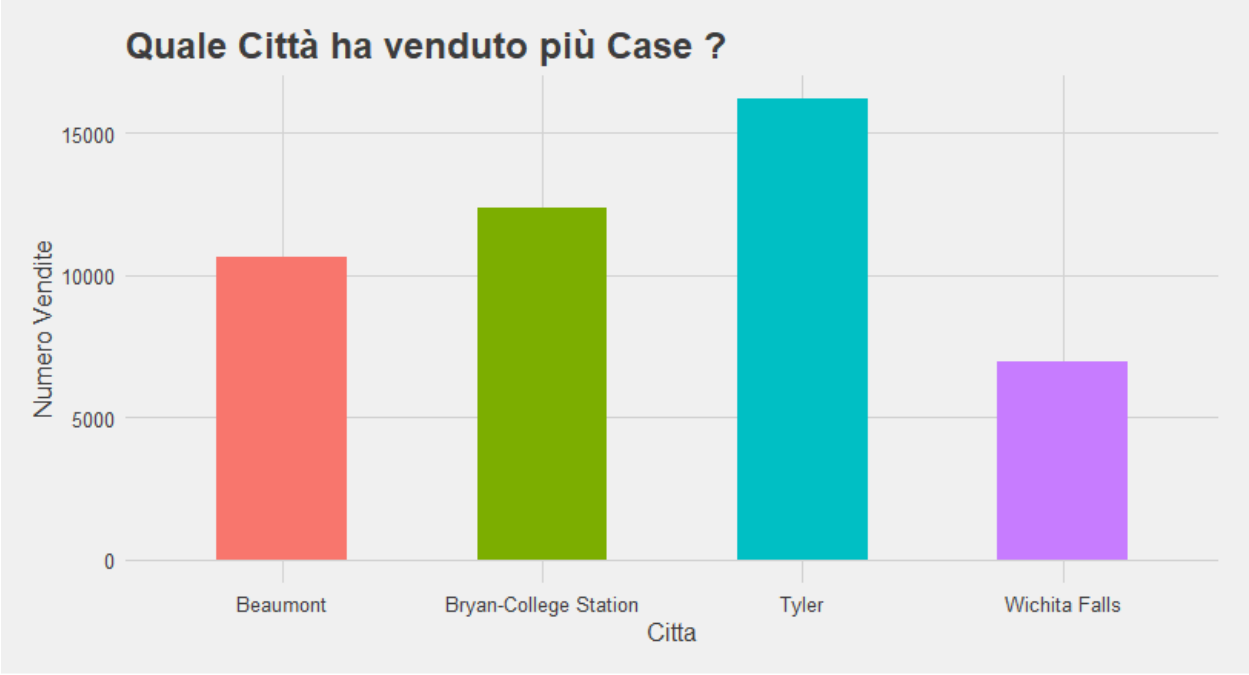


Da questo notiamo che il per il Bryan-College station ha avuto una costante crescita nella richiesta di case degli anni, che ha causato un aumento delle vendite nella zona.

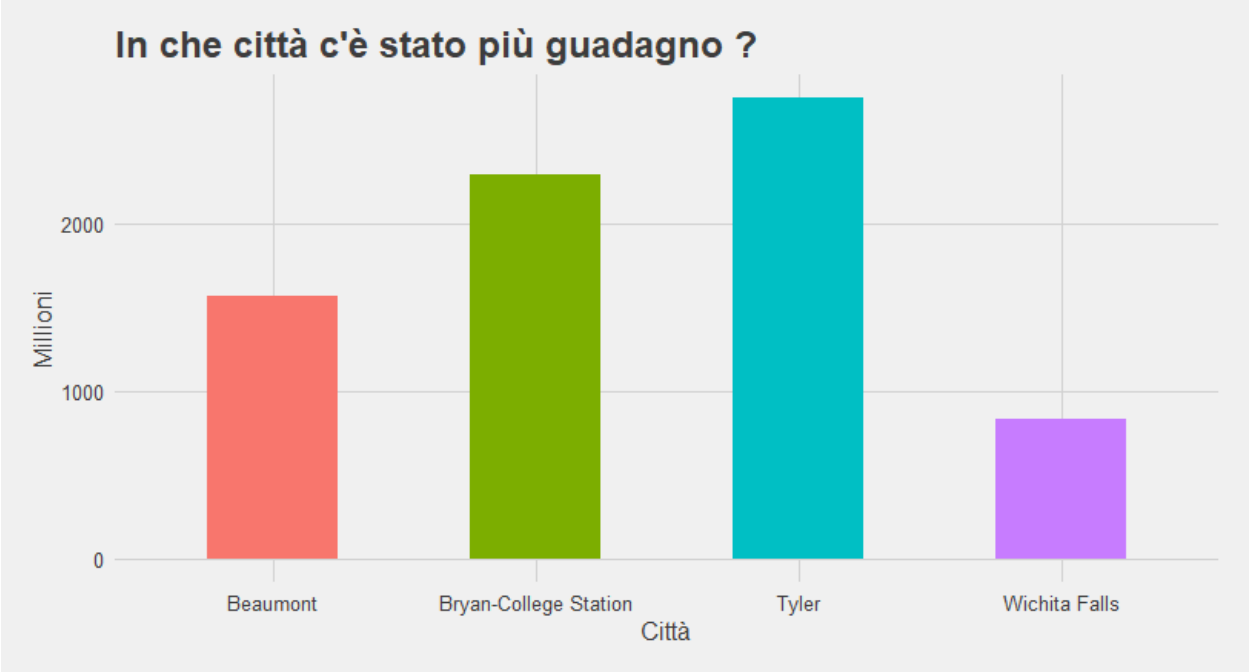
Domande

Per compiere un buona analisi statistica non basta mostrare delle variabili, osservare i dati e porsi delle domande che possano essere di supporto alle decisioni. Mi sono quindi posto una serie di domande e ho analizzato i risultati in modo critico.

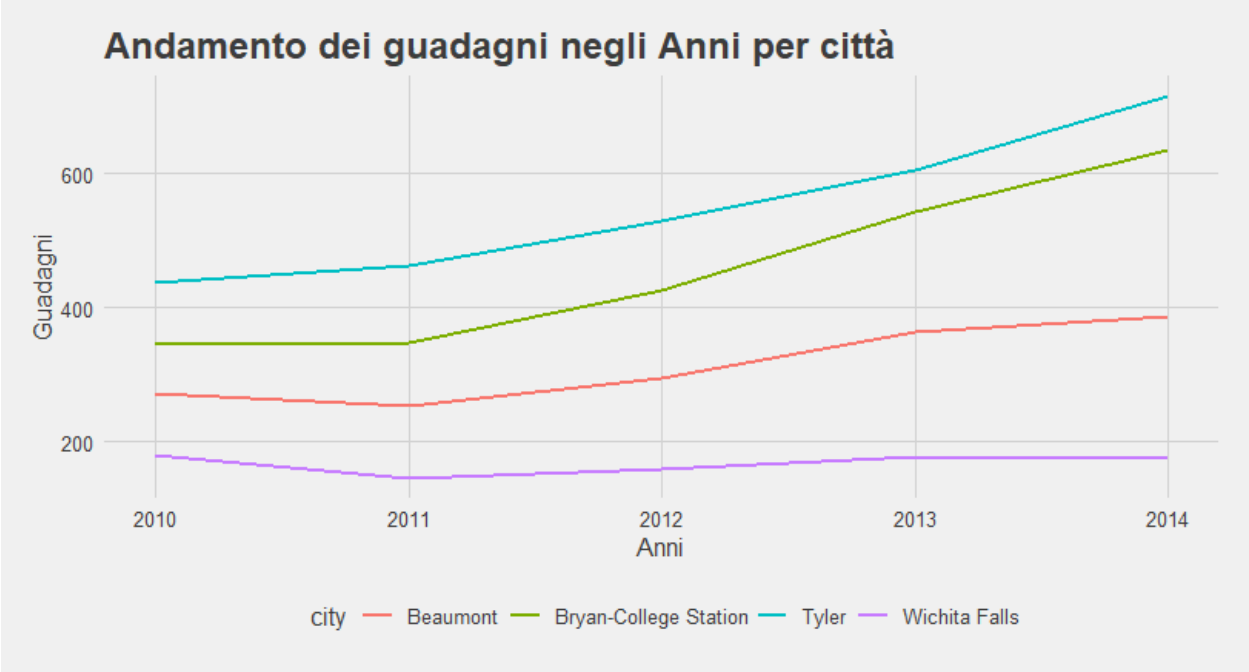
Quale la città con più vendite ?



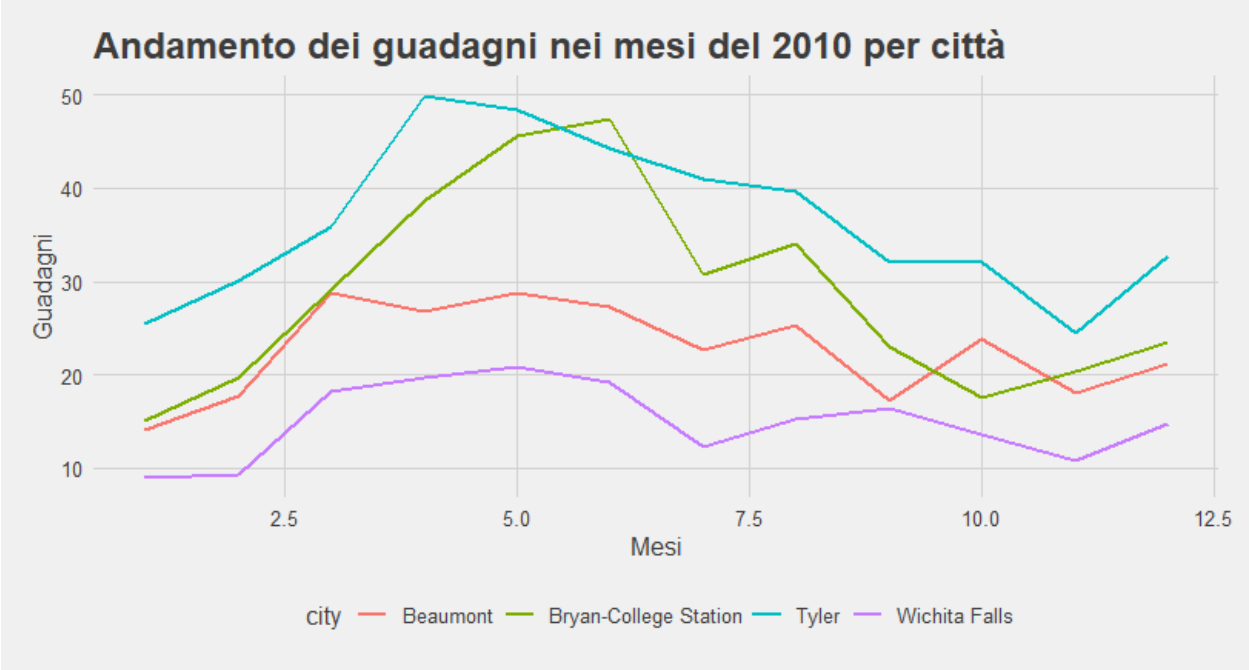
In quale città i guadagni sono stati più alti ?

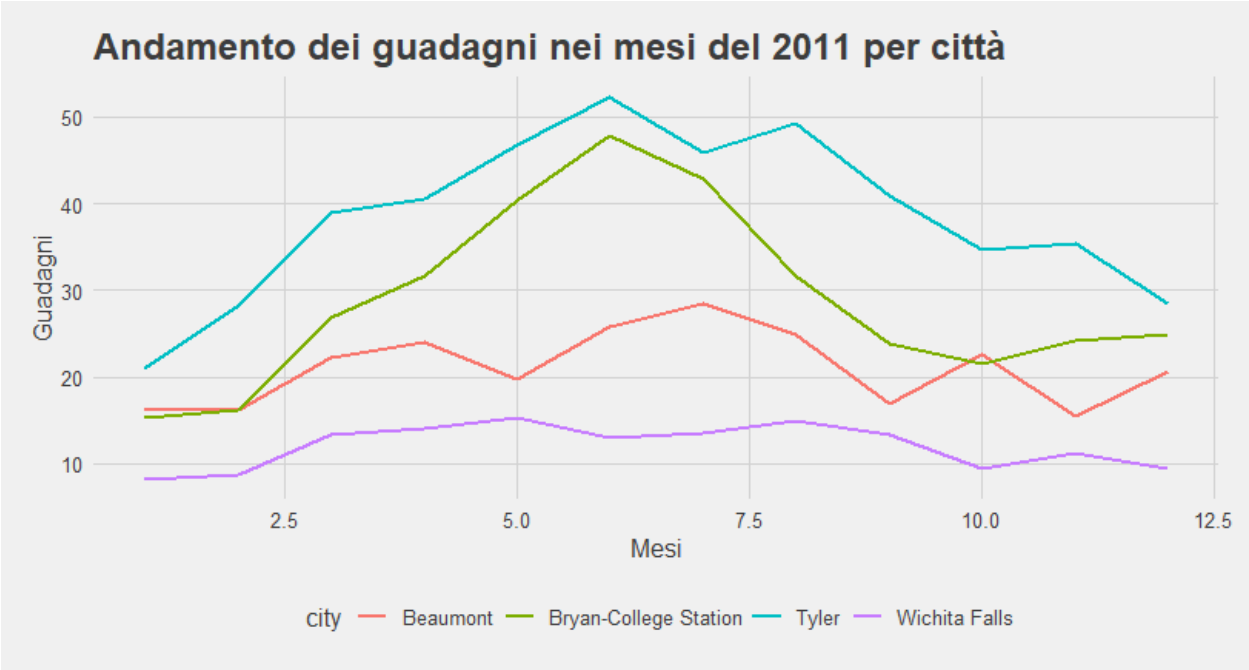


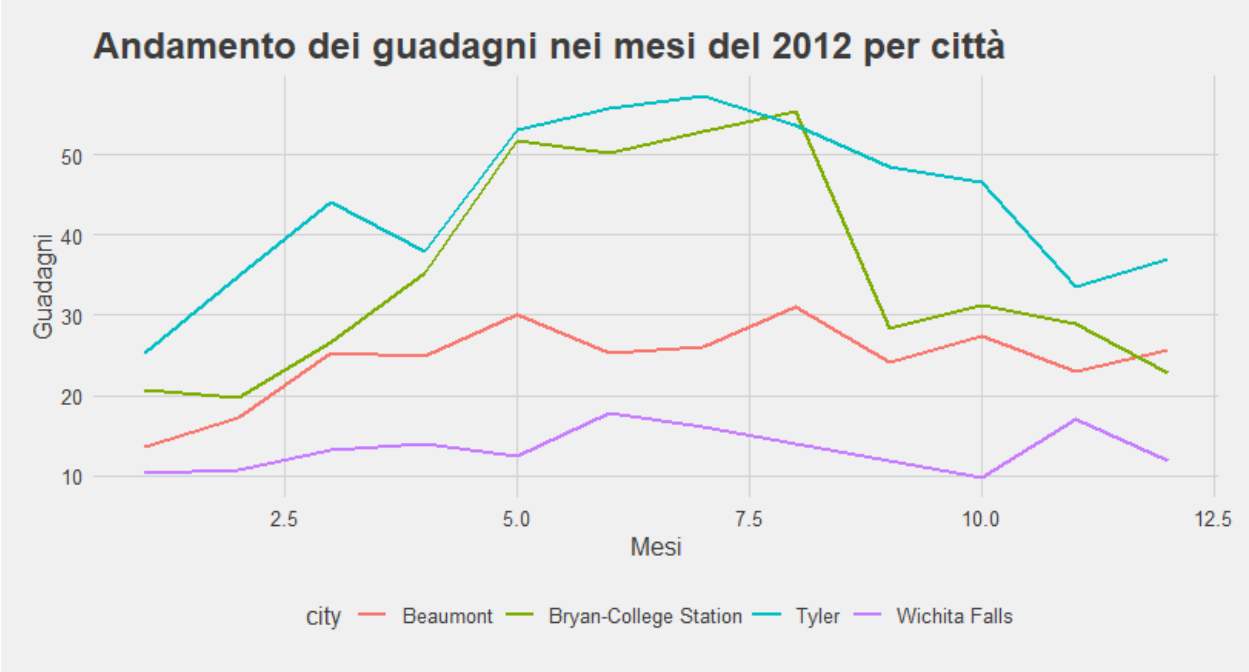
Per rendere il tutto più esplicativo mostro un grafico dell'andamento dei guadagni negli anni.

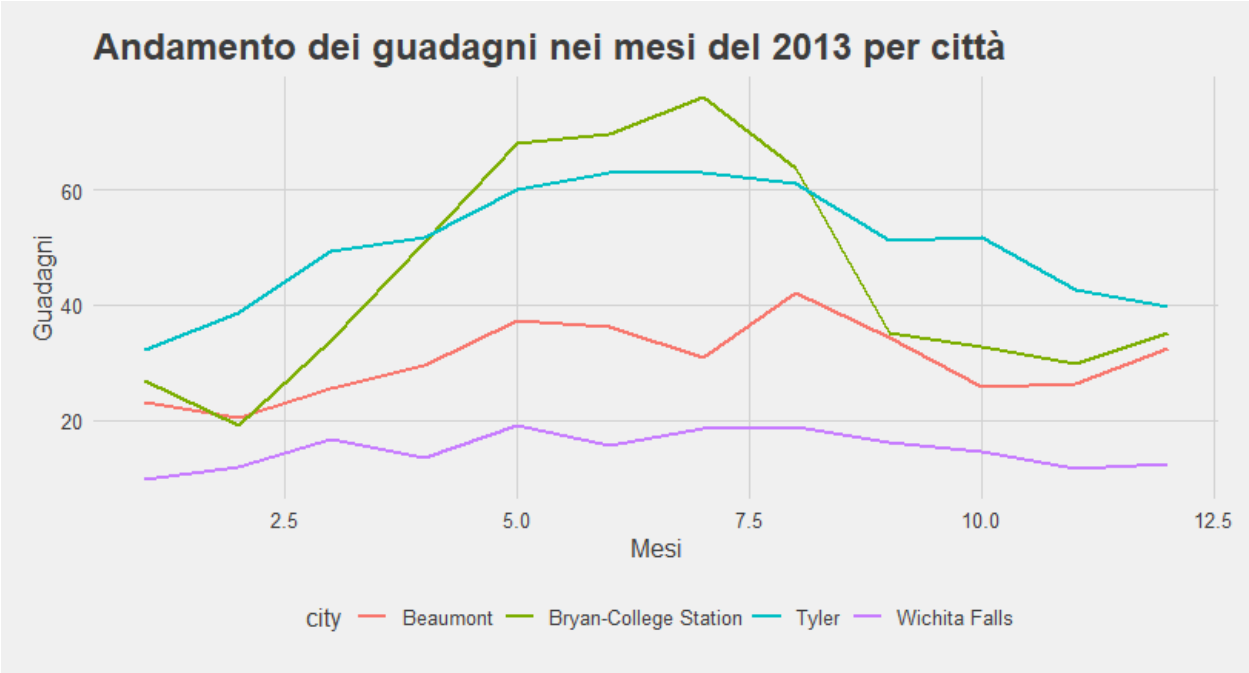


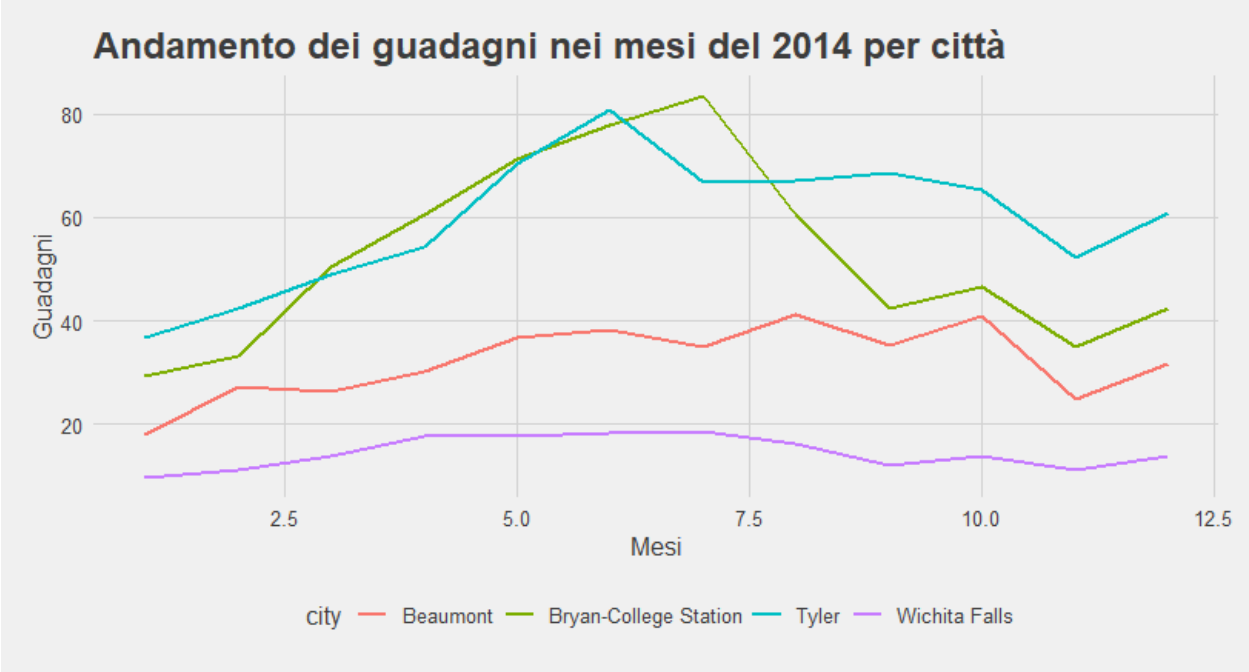
Ma andiamo ancora più a fondo, osservando il cambiamento dei guadagni nei mesi di ogni anno.


Andamento dei guadagni nei mesi del 2010 per città
Guadagni
50
40
30
20
10
2.5
5.0
7.5
10.0
12.5
Mesi
city
Beaumont
Bryan-College Station
Tyler
Wichita Falls

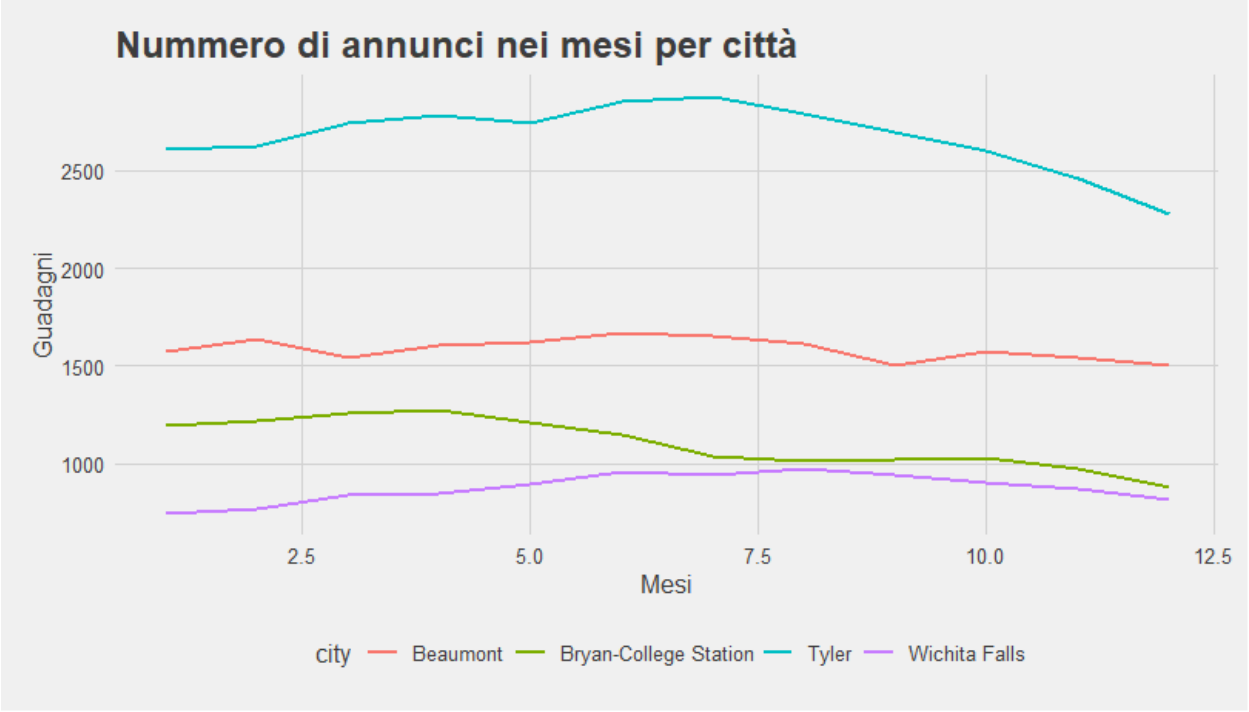

Andamento dei guadagni nei mesi del 2011 per città
Guadagni
50
40
30
20
10
2.5
5.0
7.5
10.0
12.5
Mesi
city
Beaumont
Bryan-College Station
Tyler
Wichita Falls


Andamento dei guadagni nei mesi del 2012 per città
Guadagni
50
40
30
20
10
2.5
5.0
7.5
10.0
12.5
Mesi
city
Beaumont
Bryan-College Station
Tyler
Wichita Falls


Andamento dei guadagni nei mesi del 2013 per città
Guadagni
60
40
20
2.5
5.0
7.5
10.0
12.5
Mesi
city
Beaumont
Bryan-College Station
Tyler
Wichita Falls

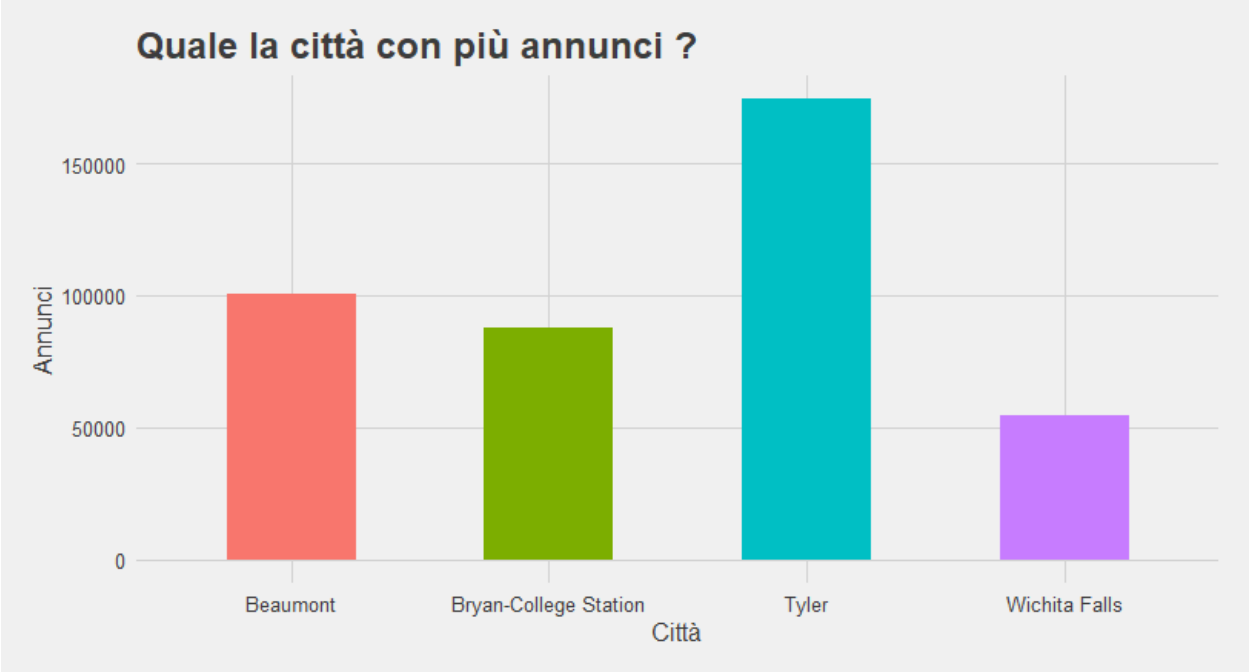


Notiamo che abbiamo un picco di guadagni sempre all'iniziò della dell'estate. Paragonando però questo grafico al'andamento del numero di annunci dello stesso anno.



Possiamo notare che il numero di annunci di vendita è stabile. Deduciamo quindi che questo picco i vendite non sia dovuto all'aumento del numero di annunci, ma ad altre cause a noi sconosciute.

In quale città ci sono più annunci di vendita ?



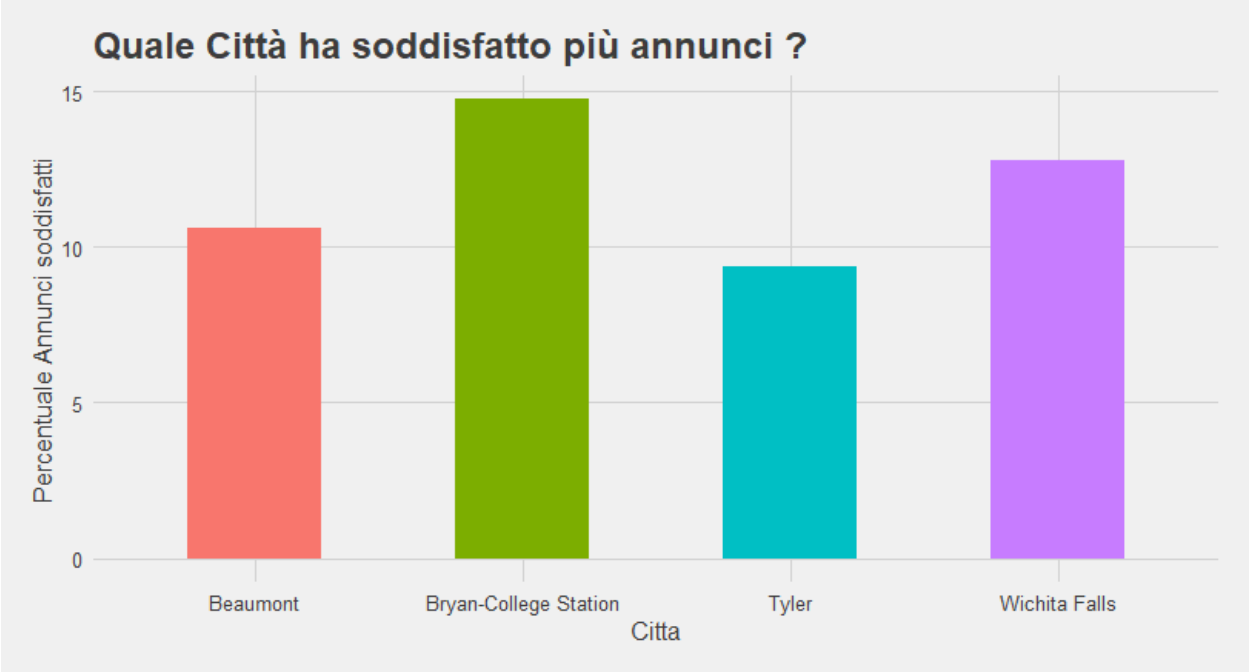
Da questo grafico scopriamo che tyler oltre a essere la città con più vendite e anche la città con più annunci di vendita. Questo ci permette di dedurre che probabilmente è un grande centro abitato che ha un alta richiesta abitativa.

Città che ha soddisfatto più annunci ?

Una colonna che potremmo aggiungere è quella relativa alla percentuale di annunci di vendita soddisfatti, come sappiamo per ogni città in un determinato anno e mese abbiamo un numero di annunci di vendite di case, abbiamo anche il numero effettivo di vendite effettuate, la nuova colonna sarà quindi la percentuale di annunci che sono riusciti a vendere casa per ogni mese. La calcoleremo (numero di vendite/numero di annunci)*100, il tutto apparirà cosi:

perc_satisfied_ads
5.414220
6.809584
10.775607
11.709602
11.405985
10.482529

Ma Osserviamo un grafico che ci mostrerà quale la città che mediamente ha soddisfatto più annunci di vendita.



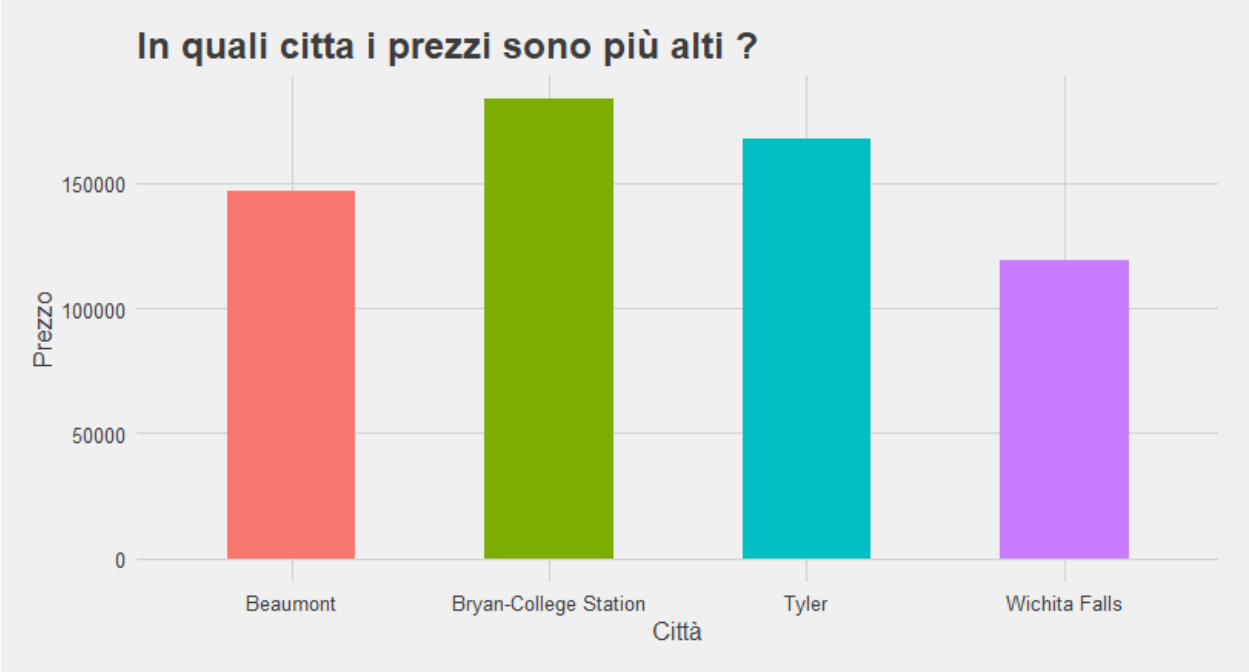
Possiamo notare che la città che ha soddisfatto più annunci di vendita e Bryan-College Station la cosa interessante e che confrontandola con gli altri grafici notiamo che è anche la città con meno annunci totali(grafico sopra) e in cui le case costavano di più (grafico sotto). Probabilmente c'è una forte richiesta dato che è un cittadina universitaria, ma proprio per questo le persone non vendono la loro casa ma magari decidono darla in affitto e quei pochi che decidono di vendere lo fanno per un prezzo più elevato.

In quale città il prezzo di vendita delle case è più alto ?

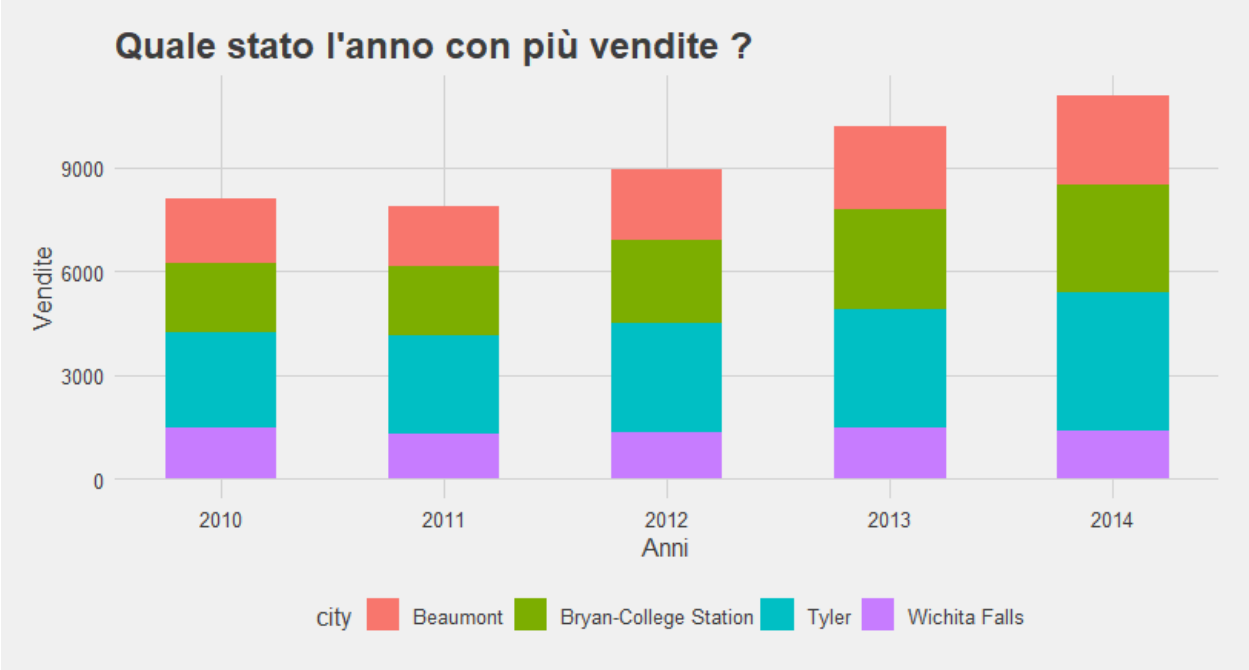
Colonna Prezzo medio
Una cosa che è utile calcolare e salvare nel dataset è il costo medio di vendita di una casa. Questo lo possiamo ottenere facilmente prendendo per ogni riga ii volume di guadagni totale dividendolo per il numero delle vendite effettuate. La colonna risulterà come segue:

	mean_price
1	170.6265
2	163.7963
3	157.6978
4	134.0950
5	142.7376
6	144.0159

Mostriamo poi attraverso un grafico il confronto tra le varie città.

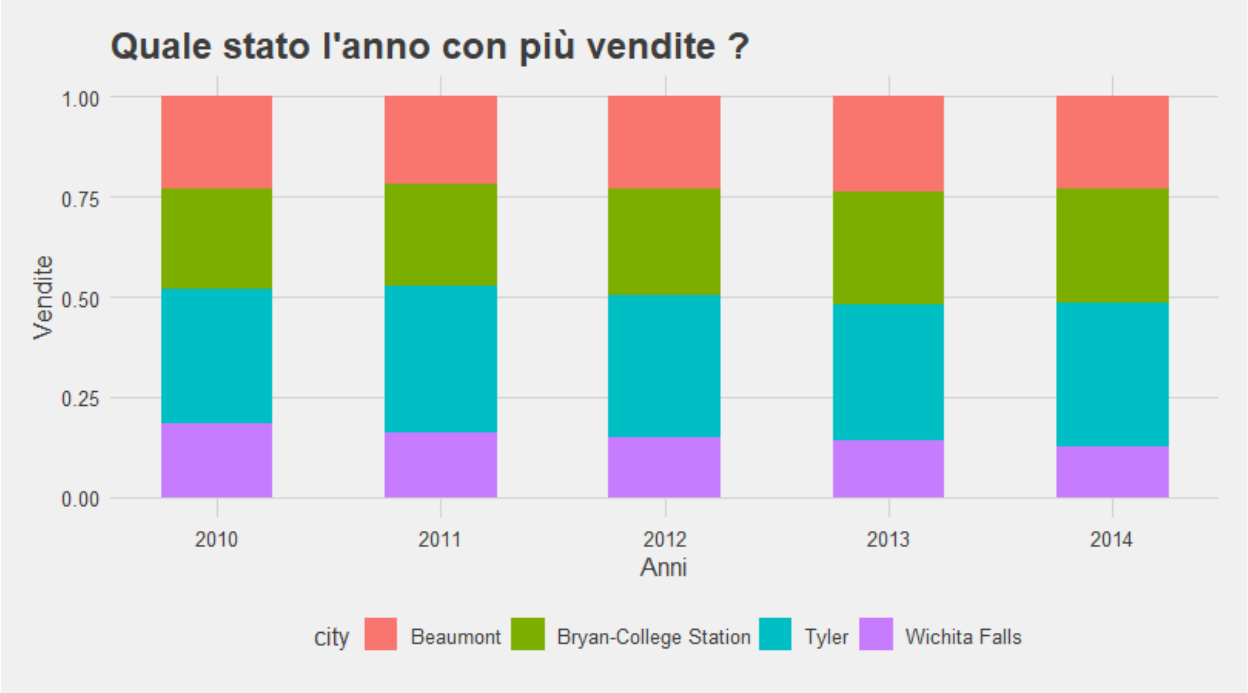


Quale l’anno in cui abbiamo venduto più case ?



Da questo grafico possiamo capire che nel tempo sta aumentando la domanda relativa alla necessita di acquistare casa. Osserviamo però delle varianti di questo grafico.

Osserviamo ora il grafico normalizzato:



Osserviamo una variante di questo grafico che paragona meglio le varie città negli anni.