



# Progetto statistica descrittiva per Data Scientist

Nome: Matteo Belletti  
Data: 09/03/2024

## Descrizione

Il progetto si propone di condurre un'analisi dettagliata sul dataset delle vendite immobiliari nella regione del Texas. Attraverso un'indagine descrittiva, esploreremo i dati relativi alle transazioni immobiliari per cercare di avere una comprensione quanto più chiara e dettagliata del mercato immobiliare texano.

## Descrizione variabili

Punto 1 della consegna			
Nome variabile	Descrizione	Tipo	Scala
city	Città	Qualitativa	Nominali
year	Anno di riferimento	Qualitativa	Ordinale
month	Mese di riferimento	Qualitativa	Nominale
sales	Numero totale di vendite	Quantitativa	Discreta
volume	Valore totale delle vendite in milioni di dollari	Quantitativa	Continua
median_price	Prezzo mediano di vendita in dollari	Quantitativa	Continua
listings	Numero totale di annunci attivi	Quantitativa	Discreta
months_inventory	Tempo necessario per vendere tutte le inserzioni correnti al ritmo attuale delle vendite, espresso in mesi	Quantitativa	Continua

## Calcolo indici

Punto 2 della consegna									
Variabile Sales									
city	mean	median	min	max	IQR	range	std	skewness_fisher	n
Beaumont	177	176	83	273	52	190	41.5	0.188	60
Bryan-College Station	206	186	89	403	148	314	85.0	0.652	60
Tyler	270	271	143	423	86.8	280	62.0	0.139	60
Wichita Falls	116	114	79	167	33	88	22.2	0.316	60
<p>1. <b>Media (mean):</b> La media aritmetica fornisce una stima del valore medio delle vendite immobiliari per città. Ad esempio, Bryan-College Station ha la media più alta (206), indicando che potrebbe avere vendite immobiliari complessivamente più elevate rispetto alle altre città.</p> <p>2. <b>Mediana (median):</b> La mediana rappresenta il valore centrale della distribuzione e non è influenzata da valori estremi. Tutte le città hanno una mediana molto simile, indicando una distribuzione relativamente simmetrica dei dati.</p> <p>3. <b>Minimo e Massimo (min e max):</b> Questi valori indicano rispettivamente il valore più basso e più alto delle vendite immobiliari per città. Ad esempio, Bryan-College Station ha il valore massimo più elevato (403), mentre Wichita Falls ha il valore minimo più basso (79).</p> <p>4. <b>Interquartile Range (IQR):</b> L'IQR rappresenta l'intervallo interquartile e fornisce una misura della dispersione dei dati intorno alla mediana. Valori più alti indicano maggiore variabilità nei dati.</p> <p>5. <b>Range:</b> Il range rappresenta la differenza tra il valore massimo e il valore minimo e offre una misura della dispersione totale dei dati. Bryan-College Station ha il range più ampio (314), indicando una maggiore variabilità nelle vendite immobiliari rispetto alle altre città.</p> <p>6. <b>Deviazione Standard (std_deviation):</b> La deviazione standard misura la dispersione dei dati intorno alla media. Valori più alti indicano maggiore variabilità.</p> <p>7. <b>Skewness di Fisher (skewness_fisher):</b> La skewness di Fisher fornisce una misura della simmetria della distribuzione dei dati. Valori vicini a zero indicano una distribuzione simmetrica.</p> <p>8. <b>Numero di Osservazioni (n):</b> Tutte le città hanno lo stesso numero di osservazioni (60), garantendo una comparabilità uniforme tra le città.</p>									

### Variabile Volume

city	mean	median	min	max	iqr	range	std_deviation	skewness_fisher	n
Beaumont	26.1	25.6	13.5	42.0	8.22	28.5	6.97	0.363	60
Bryan-College Station	38.2	33.6	15.2	83.5	23.7	68.4	17.2	0.856	60
Tyler	45.8	45.1	21.0	80.8	17.6	59.8	13.1	0.353	60
Wichita Falls	13.9	13.7	8.17	20.9	4.80	12.7	3.24	0.193	60
<p>1. <b>Media (mean):</b> Tyler ha la media più alta (45.8), suggerendo che potrebbe avere volumi di vendita più consistenti rispetto alle altre città.</p>									

2. **Mediana (median):** La mediana rappresenta il valore centrale della distribuzione e Beaumont ha la mediana più bassa (25.6), suggerendo che potrebbe avere una distribuzione dei dati leggermente più asimmetrica.
3. **Minimo e Massimo (min e max):** Questi valori indicano rispettivamente il minimo e il massimo del volume delle vendite per città. Bryan-College Station ha il valore massimo più elevato (83.5), mentre Beaumont ha il valore minimo più basso (13.5).
4. **Interquartile Range (IQR):** L'IQR rappresenta l'intervallo interquartile e Tyler ha l'IQR più alto (17.6), indicando una maggiore variabilità nei dati rispetto alle altre città.
5. **Range:** Il range è la differenza tra il valore massimo e il valore minimo del volume delle vendite. Bryan-College Station ha il range più ampio (68.4), suggerendo una maggiore variabilità nei dati.
6. **Deviazione Standard (std\_deviation):** La deviazione standard misura la dispersione dei dati intorno alla media. Tyler ha la deviazione standard più alta (13.1), indicando una maggiore variabilità nei dati.
7. **Skewness di Fisher (skewness\_fisher):** Tutte le città hanno valori prossimi a zero, indicando una distribuzione relativamente simmetrica dei dati.

#### Dati dell'Inventario

city	mean	median	min	max	iqr	range	std_deviation	skewness_fisher	n
Beaumont	9.97	10.4	7	12.6	2.83	5.6	1.65	-0.215	60
Bryan-College Station	7.66	8.1	3.4	11.6	4	8.2	2.25	-0.315	60
Tyler	11.3	11.4	6.9	14.9	2.77	8	1.89	-0.110	60
Wichita Falls	7.82	7.9	6.1	9.4	1.03	3.3	0.781	-0.226	60

1. **Media (mean):** Tyler ha la media più alta (11.3), suggerendo che potrebbe avere più annunci attivi rispetto alle altre città.
2. **Mediana (median):** Tutte le città hanno mediana molto simili, suggerendo una distribuzione relativamente simmetrica dei dati.
3. **Minimo e Massimo (min e max):** Beaumont ha il valore massimo più elevato (12.6), mentre Wichita Falls ha il valore minimo più basso (6.1).
4. **Interquartile Range (IQR):** Tyler ha l'IQR più alto (2.77), indicando una maggiore variabilità nei dati rispetto alle altre città.
5. **Range:** Bryan-College Station ha il range più ampio (8.2), suggerendo una maggiore variabilità nei dati.
6. **Deviazione Standard (std\_deviation):** Tyler ha la deviazione standard più alta (1.89), indicando una maggiore variabilità nei dati.
7. **Skewness di Fisher (skewness\_fisher):** Tutte le città hanno valori prossimi a zero, indicando una distribuzione relativamente simmetrica dei dati.

#### Variabile listing

city	mean	median	min	max	iqr	range	std_deviation	skewness_fisher	n
Beaumont	1679.0	1676	1500	1857	124.	357	91.1	0.0195	60
Bryan-College Station	1458.0	1489	882	1984	350.	1102	253.0	-0.345	60
Tyler	2905.0	2888	2272	3296	303.	1024	227.0	-0.163	60
Wichita Falls	910.0	911	743	1052	104.	309	73.8	-0.250	60

#### Considerazioni:

1. **Media (mean):** La media del numero di annunci attivi fornisce un'indicazione del numero medio di annunci attivi per città. Tyler ha la media più alta (2905), suggerendo che potrebbe avere un mercato immobiliare più dinamico rispetto alle altre città.
2. **Mediana (median):** La mediana rappresenta il valore centrale della distribuzione e tutte le città hanno una mediana molto simile, suggerendo una distribuzione relativamente simmetrica dei dati.
3. **Minimo e Massimo (min e max):** Questi valori indicano rispettivamente il minimo e il massimo del numero di annunci attivi per città. Bryan-College Station ha il valore massimo più elevato (1984), mentre Wichita Falls ha il valore minimo più basso (743).
4. **Interquartile Range (IQR):** L'IQR rappresenta l'intervallo interquartile e Tyler ha l'IQR più alto (303), indicando una maggiore variabilità nei dati rispetto alle altre città.
5. **Range:** Il range è la differenza tra il valore massimo e il valore minimo del numero di annunci attivi. Bryan-College Station ha il range più ampio (1102), suggerendo una maggiore variabilità nei dati.
6. **Deviazione Standard (std\_deviation):** Tyler ha la deviazione standard più alta (227.0), indicando una maggiore variabilità nei dati.
7. **Skewness di Fisher (skewness\_fisher):** La skewness di Fisher fornisce una misura della simmetria della distribuzione dei dati. Tutte le città hanno valori prossimi a zero, indicando una distribuzione relativamente simmetrica dei dati.

#### Dati del Prezzo Mediano di Vendita:


city	mean	median	min	max	iqr	range	std_deviation	skewness_fisher	n
Beaumont	129988	130750	106700	163800	11525	57100	10105	0.362	60
Bryan-College Station	157488	155400	140700	180000	11175	39300	8852	0.713	60
Tyler	141442	142200	120600	161600	13700	41000	9337	0.124	60
Wichita Falls	101743	102300	73800	135300	16375	61500	11320	0.215	60

#### Considerazioni:

1. **Media (mean):** La media del prezzo medio di vendita indica il valore medio delle vendite immobiliari per città. Bryan-College Station ha la media più alta (157488), suggerendo che potrebbe avere prezzi medi di vendita più elevati rispetto alle altre città.
2. **Mediana (median):** La mediana rappresenta il valore centrale della distribuzione e tutte le città hanno valori molto simili, suggerendo una distribuzione relativamente simmetrica dei dati.
3. **Minimo e Massimo (min e max):** Questi valori indicano rispettivamente il minimo e il massimo del prezzo medio di vendita per città. Beaumont ha il valore massimo più elevato (163800), mentre Wichita Falls ha il valore minimo più basso (73800).
4. **Interquartile Range (IQR):** L'IQR rappresenta l'intervallo interquartile e tutte le città hanno valori abbastanza simili, indicando una distribuzione relativamente uniforme dei dati.
5. **Range:** Il range è la differenza tra il valore massimo e il valore minimo del prezzo medio di vendita. Beaumont ha il range più ampio (57100), suggerendo una maggiore variabilità nei dati.

6. **Deviazione Standard (std\_deviation):** Beaumont ha la deviazione standard più alta (10105), indicando una maggiore variabilità nei dati.
7. **Skewness di Fisher (skewness\_fisher):** Tutte le città hanno valori prossimi a zero, indicando una distribuzione relativamente simmetrica dei dati.

### Identificare la variabile con la massima variabilità

 Punto 3 della consegna

Questi gli indici non divisi per città:

**Sales Summary:**

mean	median	min	max	iqr	range	std_deviation	skewness_fisher	cv	n
192.2917	175.5	79	423	120	344	79.65111	0.718104	0.4142203	240

**Volume Summary:**

mean	median	min	max	iqr	range	std_deviation	skewness_fisher	cv	n
31.00519	27.0625	8.166	83.547	23.2335	75.381	16.65145	0.884742	0.5370536	240

**Median Price Summary:**

mean	median	min	max	iqr	range	std_deviation	skewness_fisher	cv	n
132665.4	134500	73800	180000	32750	106200	22662.15	-0.3645529	0.1708218	240

**Listings Summary:**

mean	median	min	max	iqr	range	std_deviation	skewness_fisher	cv	n
1738.021	1618.5	743	3296	1029.5	2553	752.7078	0.6494982	0.4330833	240

**Months Inventory Summary:**


mean	median	min	max	iqr	range	std_deviation	skewness_fisher	cv	n
9.1925	8.95	3.4	14.9	3.15	11.5	2.303669	0.04097527	0.2506031	240

Per determinare quale variabile presenta la variabilità maggiore, è necessario utilizzare metriche che non siano influenzate (o lo siano in misura minore) dalla scala di misurazione dei dati. In questo caso, possiamo considerare due indici:

- **Skewness di Fisher (skewness\_fisher):** Misura l'asimmetria della distribuzione di una variabile.
- **Coefficiente di variazione (CV):** Rappresenta il rapporto tra la deviazione standard e la media, fornendo una misura di variabilità relativa.

Analizzando entrambi gli indici, si osserva che **Volume** presenta i valori maggiori. Di conseguenza, possiamo affermare che **Volume** è la variabile con la maggiore variabilità all'interno del dataset.

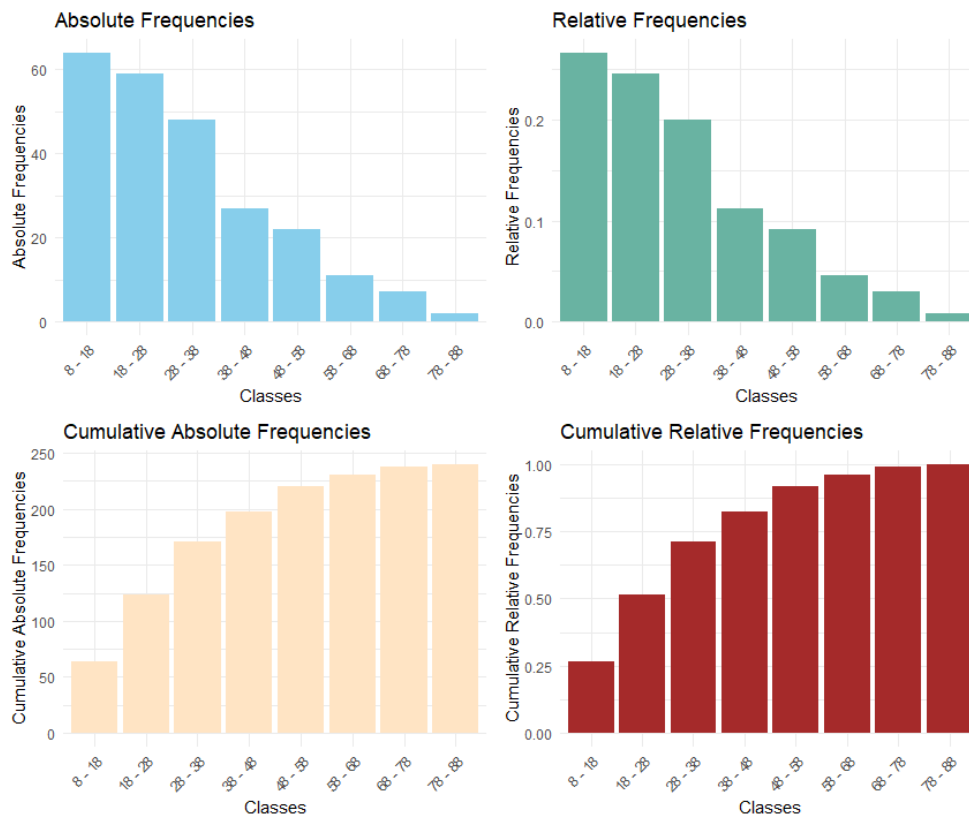
### Divisione variabile Volume per classi

 Punto 4 della consegna

**Tabella**

volume_range	absolute_freq	relative_freq	cumulative_absolute_freq	cumulative_relative_freq
8 - 18	64	0.266666667	64	0.2666667
18 - 28	59	0.245833333	123	0.5125000
28 - 38	48	0.200000000	171	0.7125000
38 - 48	27	0.112500000	198	0.8250000
48 - 58	22	0.091666667	220	0.9166667
58 - 68	11	0.045833333	231	0.9625000
68 - 78	7	0.029166667	238	0.9916667
78 - 88	2	0.008333333	240	1.0000000

**Grafici**



L'indice di Gini risulta essere: 0.2957647

### Indice di Gini per la variabile City

💡 Punto 5 della consegna

Il dataset contiene 60 osservazioni per ogni città, l'indice di Gini per la variabile city è quindi:

| 0.25

### Calcolo probabilità

💡 Punto 6 della consegna

#### Calcolo probabilità Luglio

Il dataset contiene 240 osservazioni divise equamente nell'arco di 5 anni per 4 città, il mese di luglio apparirà quindi

|  $4 * 5 = 20$

volte, la probabilità quindi di pescare a caso il mese di Luglio è

|  $20 / 240 = 0,083$  (8,33 %)

Pari ad ogni altro mese dell'anno

#### Calcolo probabilità Dicembre 2012

La probabilità di pescare a caso uno delle 4 righe che riportano Dicembre 2012 sulle 240 osservazioni è:

$4/240 = 0.016666$  (1,67%)

#### Calcolo probabilità città Beaumont

Il dataset contiene 240 osservazioni divise equamente nell'arco di 5 anni per 4 città, il numero delle osservazioni per ogni città è

|  $240 / 4 = 60$

volte, la probabilità quindi di pescare a caso la città Beaumont (o le altre...) è:

|  $60 / 240 = 0,25$  (25 %)

## Prezzo Medio

### Punto 7 della consegna

Per questo questi è stato sviluppato il metodo R:

```
# Define a function to get the mean price from a dataset
get_mean_price <- function(df){
  # Calculate the mean price
  summary <- df %>%
    summarise(mean_price = sum(volume) / sum(sales) * 1000)

  return(summary) # Return the summary
}
```

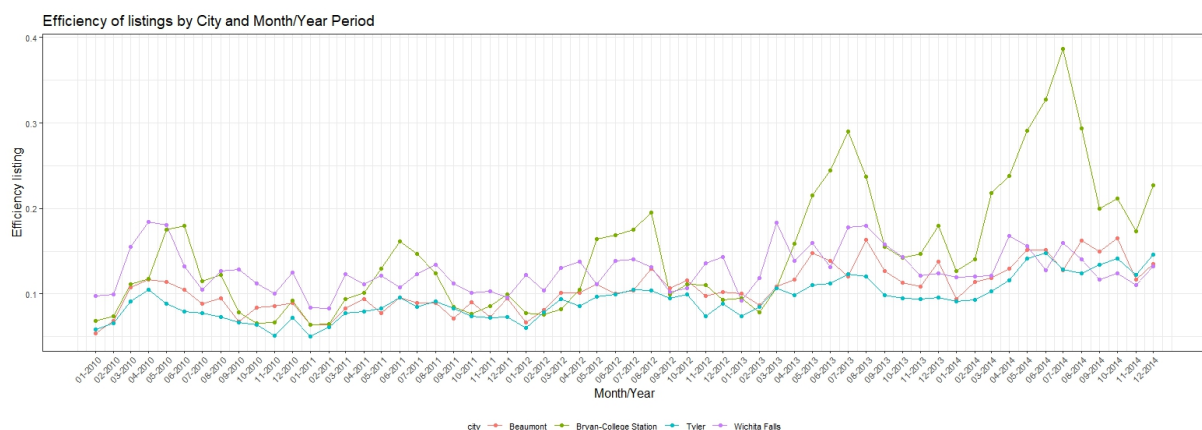
Che al dataset passato come parametro aggiunge una colonna con il prezzo medio, in modo che possa essere riutilizzato nei vari raggruppamenti che si desidera.

## Efficiency delle vendite

### Punto 8 della consegna

Come indice per calcolare l'efficacia degli annunci di vendita è stato scelto il rapporto tra vendite e annunci attivi della riga.

I risultati sono rappresentati dal seguente grafico



Dal grafico emerge che la città di Bryan-College Station presenta dei picchi nei valori che corrispondono a periodi di aumento delle vendite mensili, mantenendo il numero di annunci di vendita nello stesso ordine di grandezza.

## Summary

### Punto 9 della consegna

Nel progetto la generazione dei summary è all'interno dell'area marcata con il codice:

```
#####
# Point 10
# summary data for city and month
#####
```

Inoltre è stata usata sempre la medesima funzione utilizzata per analizzare i punti precedenti, un esempio di output è il seguente.

### Risultati per la variabile "listings" per la città Beaumont

city	month	mean	geom_mean	median	min	max	iqr	range	std_deviation	skewness_fit
Beaumont	1	1603.	1602.	1581	1533	1677	72	144	58.3	0.185
Beaumont	2	1640.	1639.	1636	1586	1691	46	105	40.6	-0.0472
Beaumont	3	1657.	1655.	1689	1539	1762	163	223	99.2	-0.240
Beaumont	4	1702	1701.	1708	1604	1767	39	163	61.2	-0.768
Beaumont	5	1729.	1728.	1765	1620	1832	112	212	87.2	-0.194
Beaumont	6	1744.	1742.	1724	1672	1845	128	173	77.5	0.324
Beaumont	7	1759.	1757.	1749	1657	1857	114	200	81.7	0.0146

city	month	mean	geom_mean	median	min	max	iqr	range	std_deviation	skewness_fit
Beaumont	8	1719.	1717.	1683	1617	1830	114	213	87.8	0.221
Beaumont	9	1700	1696.	1704	1501	1829	104	328	126.	-0.711
Beaumont	10	1680.	1679.	1671	1575	1779	67	204	76.3	-0.0991
Beaumont	11	1650.	1648.	1652	1544	1742	63	198	73.7	-0.255
Beaumont	12	1569.	1568.	1570	1500	1646	62	146	56.2	0.153

#### Risultati per la variabile "listings" per la città Wichita Falls

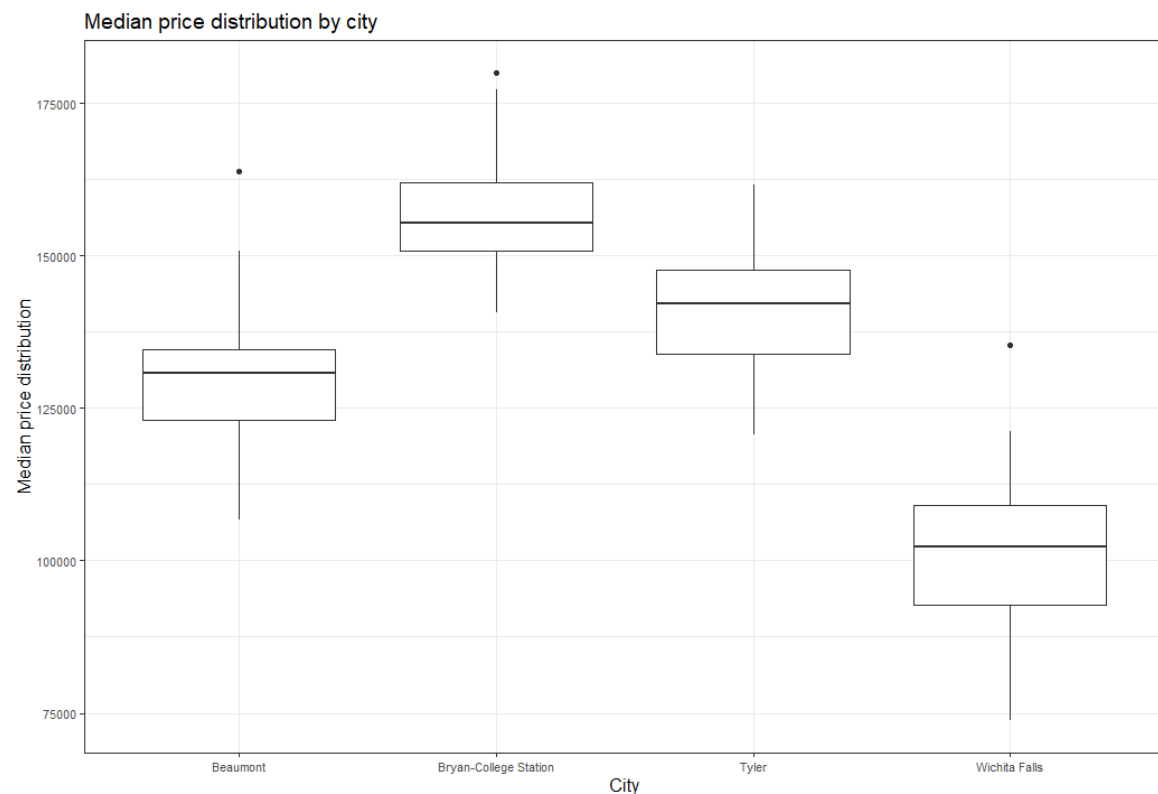
city	month	mean	geom_mean	median	min	max	iqr	range	std_deviation	skewness_fit
Wichita Falls	1	864.	862.	859	746	955	54	209	77.9	-0.494
Wichita Falls	2	870	868.	861	774	950	65	176	67.3	-0.266
Wichita Falls	3	901.	900.	887	838	968	78	130	54.2	0.156
Wichita Falls	4	911.	910.	904	852	996	12	144	52.5	0.793
Wichita Falls	5	936.	934.	914	899	1052	14	153	65.4	1.45
Wichita Falls	6	964	963.	961	923	1030	38	107	41.9	0.732
Wichita Falls	7	950.	947.	941	844	1029	52	185	69.8	-0.490
Wichita Falls	8	952.	950.	973	830	1022	71	192	76.3	-0.854
Wichita Falls	9	943.	940.	940	812	1028	74	216	84.2	-0.661
Wichita Falls	10	915.	912.	907	796	1005	58	209	78.6	-0.496
Wichita Falls	11	879.	877.	877	777	968	32	191	68.8	-0.286
Wichita Falls	12	829.	827.	821	743	938	43	195	71.3	0.474

### Boxplot prezzo mediano



Punto 1 della parte di grafici della consegna

### Grafico



### Considerazioni

Bryan-College Station emerge come la città con il prezzo medio più elevato, seguita da Tyler. Questo suggerisce che queste due città potrebbero avere un mercato immobiliare più costoso rispetto alle altre considerate nel dataset.

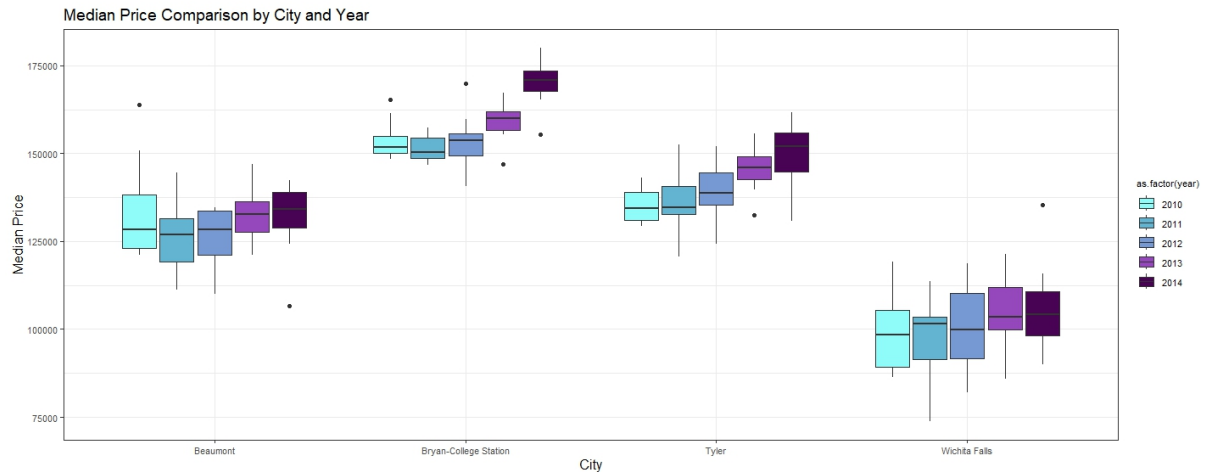
Al contrario, Wichita Falls si distingue per avere una media di valori più bassi rispetto alle altre città analizzate. Tuttavia, ciò che risulta particolarmente interessante è la sua maggiore dispersione dei dati. Questo è evidenziato sia dall'Interquartile Range (IQR), sia dalla presenza di valori outsider, cioè quei valori che si discostano significativamente dalla media.

Questo suggerisce che mentre i prezzi medi possono essere più bassi a Wichita Falls, c'è una gamma più ampia di prezzi che possono essere osservati, indicando una maggiore eterogeneità nel mercato immobiliare.

### Boxplot prezzo mediano per anno

💡 Punto 2 della parte di grafici della consegna

## Grafico



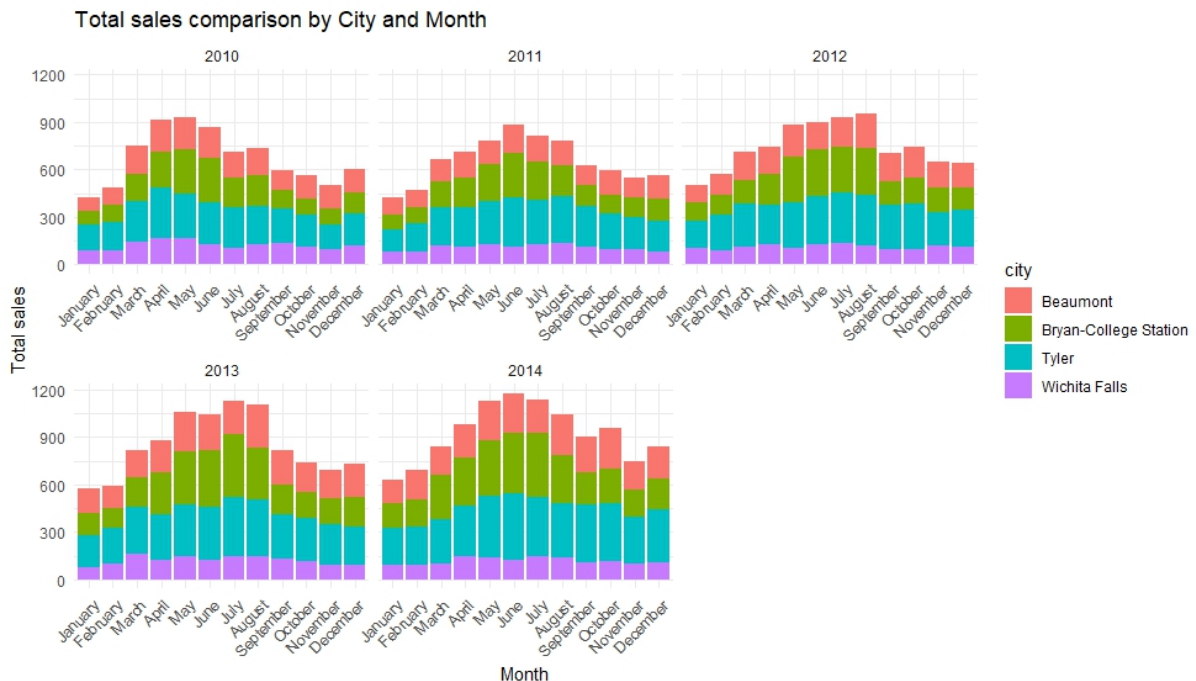
## Considerazioni

- Il grafico conferma che la città di Bryan-College Station ha consistentemente valori più elevati rispetto alle altre città nel corso degli anni, e presenta anche una minore dispersione dei dati. Inoltre, Bryan-College Station ha registrato un aumento maggiore del prezzo mediano rispetto alle altre città nel periodo considerato.
- Per quanto riguarda Wichita Falls, le osservazioni rimangono valide: la città presenta valori medi inferiori e una maggiore dispersione dei dati rispetto alle altre città. Tuttavia, nel corso degli anni, questa differenza di dispersione sembra essere diminuita, con i valori più recenti per l'anno 2014 che sono più simili a quelli di Tyler e Beaumont.
- Le città di Beaumont e Wichita Falls hanno mostrato nel tempo un incremento dei prezzi più contenuto rispetto a Tyler e soprattutto rispetto a Bryan-College Station.

## Vendite nel corso dei mesi

💡 Punto 3 della parte di grafici della consegna

## Grafico non normalizzato



## Considerazioni

- Picco delle vendite annuali:** Nel corso degli anni, il picco delle vendite si è spostato da Aprile/Giugno a Maggio/Agosto, indicando un cambiamento nei modelli di acquisto nel tempo.

2. **Minimi stagionali:** Gennaio e Febbraio mantengono costantemente i minimi delle vendite, suggerendo una stagionalità nel mercato immobiliare.
3. **Decremento significativo a settembre:** Un forte calo delle vendite si osserva nel mese di settembre.

## Grafico normalizzato



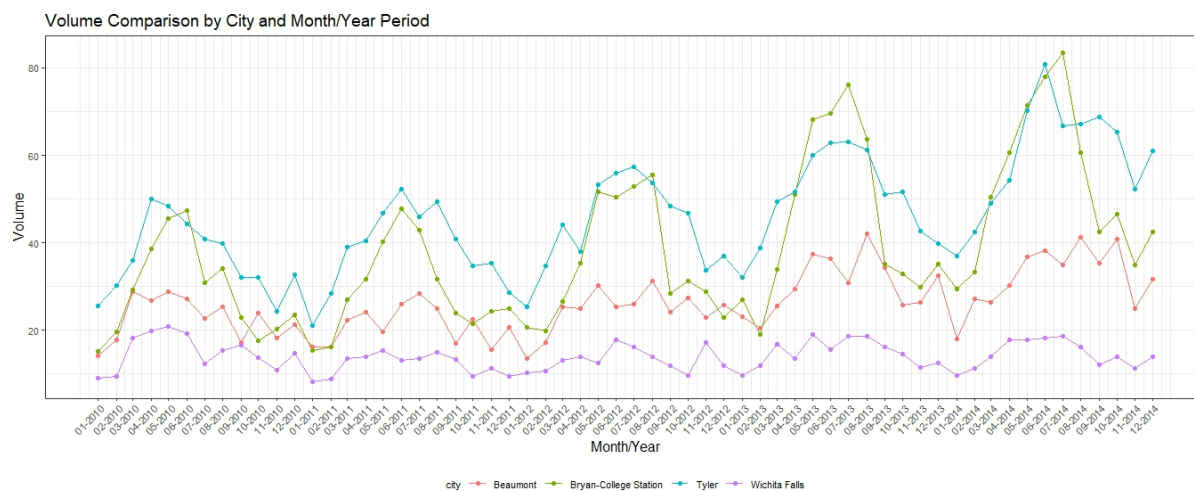
## Considerazioni

1. **Prevalenza delle vendite a Tyler:** Tyler rappresenta la maggior parte delle vendite immobiliari, costituendo almeno il 35% del totale in molti mesi e anni.
2. **Riduzione del divario con Bryan-College Station:** Nei mesi da maggio ad agosto, tranne che nell'anno 2010, il divario percentuale tra Tyler e Bryan-College Station si riduce, con Bryan-College Station che talvolta supera Tyler.
3. **Variazioni percentuali:** Le percentuali di vendite di Bryan-College Station mostrano maggiore variabilità rispetto ad altre città, che tendono ad avere comportamenti più costanti nel tempo.

## Vendite nel corso dei mesi

📌 Punto 3 della parte di grafici della consegna

## Grafico





## Considerazioni

1. **Primato di Tyler nel volume delle vendite:** Tyler conferma il suo primato anche nel volume delle vendite nella maggior parte dei periodi analizzati.
2. **Differenze di prezzo e volume:** Il grafico del volume delle vendite mostra gap meno accentuati rispetto al grafico delle vendite. Ciò suggerisce che, nonostante Bryan-College Station abbia un prezzo mediano mediamente più elevato, potrebbe registrare meno vendite ma di valore più alto.
3. **Stagionalità delle vendite:** Tyler e Bryan-College Station mostrano una sorta di stagionalità nei volumi di vendita, mentre Beaumont e Wichita sembrano meno influenzate dal cambiamento stagionale.
4. **Incremento del prezzo e volume:** Nonostante l'incremento del prezzo mediano di vendita a Wichita, non si osserva un corrispondente aumento nel volume delle vendite, suggerendo dinamiche diverse rispetto ad altre città.