Analisi mercato immobiliare Texas

1) Ho iniziato importando il dataset attraverso il comando “read.csv”.

2) Nel dataset abbiamo otto variabili di diverso tipo. In particolare:
   -CITY: QUALITATIVA su scala NOMINALE.
   -YEAR e MONTH: anche se sono rappresentate da valori numerici in quest'analisi le andremo a considerare come ORDERED FACTOR, vengono infatti utilizzate per indicare determinati periodi ordinati temporalmente.
   -SALES, MEDIAN_PRICE, LISTINGS: QUANTITATIVE DISCRETE, assumono solo valori interi. -VOLUME e MONTHS_INVENTORY: QUANTITATIVE CONTINUE, assumono solo valori float.

3) Vado quindi a calcolare i vari indici di posizione, variabilità e forma per tutte le variabili per il quale ha senso farlo. I risultati fanno riferimento alla somma dei dati di ogni città. Ogni analisi sarà accompagnata da un breve commento sui dati ottenuti.

3.1) Sales:

mediana	sd	asimmetria	curtosi
175.5	79.65111	0.718104	-0.3131764

In media vengono vendute 175 case al mese, con una deviazione standard di circa 80 vendite. Abbiamo una distribuzione asimmetrica positiva platicurtica.

3.2) Volume:

mediana	sd	asimmetria	curtosi
27.0625	16.65145	0.884742	0.176987

In media abbiamo un valore totale delle vendite mensili pari a 27 milioni di dollari, con una deviazione standard di 16.65 milioni di dollari. La distribuzione è asimmetrica positiva leptocurtica.

3.3) Median_price:

mediana	sd	asimmetria	curtosi
134500	22662.15	-0.3645529	-0.6229618

Il prezzo medio di vendita di una casa è pari a 134500 dollari, con una deviazione standard di 22662 dollari. La distribuzione è asimmetrica negativa platicurtica.

3.4) Listings:

mediana	sd	asimmetria	curtosi
1618.5	752.7078	0.6494982	-0.79179

Ogni mese ci sono in media 1618 annunci attivi, con una deviazione standard di 753 annunci. La distribuzione è asimmetrica positiva platicurtica.

3.5) Months_inventory:

mediana	sd	asimmetria	curtosi
8.95	2.303669	0.04097527	-0.1744475

Per vendere tutte le inserzioni servirebbero in media 9 mesi, con una deviazione standard di circa 2 mesi. La distribuzione è lievemente asimmetrica positiva platicurtica.

4) A questo punto vorremmo sapere qual è la variabile con variabilità più elevata e quella più asimmetrica. Avendo a che fare con variabili aventi differenti unità di misura conviene calcolare il coefficiente di variazione di ogni variabile.

sales_cv	volume_cv	median_price_cv	listings_cv	months_inventory_cv
41.42203	53.70536	17.08218	43.30833	25.06031

Dai risultati ottenuti possiamo notare che la “variabile più variabile” è il volume, ossia il valore totale delle vendite in milioni di dollari, con 53.70 punti percentuali.
Passando all’asimmetria:

sales_asimmetria	volume_asimmetria	median_price_asimmetria
0.718104	0.884742	-0.3645529

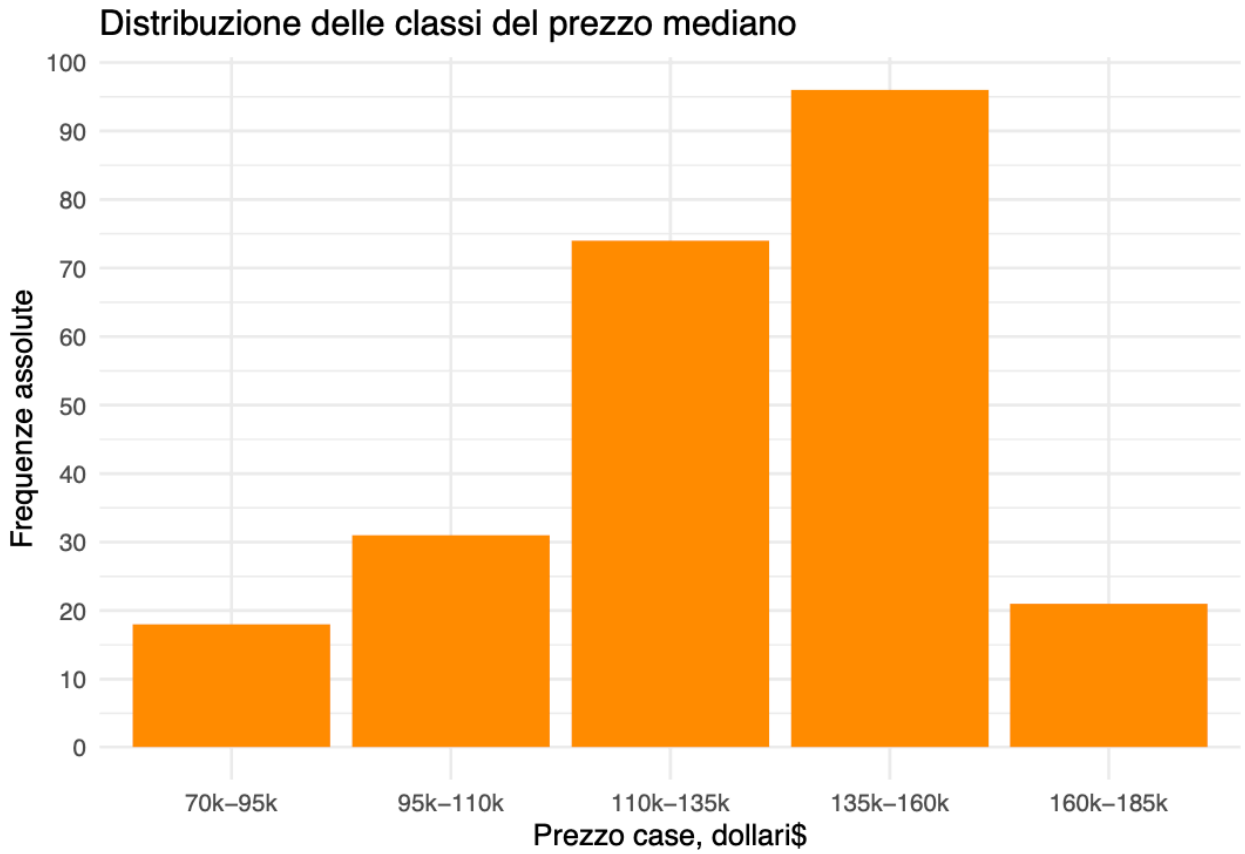
listings_asimmetria	months_inventory_asimmetria
0.6494982	0.04097527

Prendendo il valore assoluto di ogni risultato calcolato possiamo notare che, come conseguenza della sua elevata variabilità, volume è anche la variabile più asimmetrica.

5) Arrivato a questo punto ho deciso di dividere la variabile “median_price” in classi. Ho calcolato min e MAX, rispettivamente di 73800 e 180000, e creato sei classi che vanno da 70000 a 185000 con un intervallo di 25000.
Ho costruito poi la tabella con le distribuzioni di frequenze:

	ni	fi	Ni	Fi
70k-95k	18	0.0750000	18	0.0750000
95k-110k	31	0.1291667	49	0.2041667
110k-135k	74	0.3083333	123	0.5125000
135k-160k	96	0.4000000	219	0.9125000
160k-185k	21	0.0875000	240	1.0000000

Ed ho creato un grafico a barre con le frequenze assolute:



Dal grafico notiamo che la classe più popolata è quella con range di prezzo 135k-160k.

Ho infine calcolato l’indice di Gini, che può assumere valori che vanno da 0 ad 1, dove 0 sta per omogeneità e 1 per eterogeneità massima. Con un indice pari a 0.89 ci troviamo di fronte ad una variabile con un’alta eterogeneità.

6) Seguendo il ragionamento fatto prima, se andiamo a calcolare l'indice di Gini per la variabile CITY notiamo che con un indice pari ad 1 CITY ha eterogeneità massima, è quindi equidistribuita.

7) A Andiamo quindi a calcolare un po' di probabilità. Presa una riga a caso:
   -La probabilità che essa riporti la città "Beaumont" è di 60/240, cioè il 25%.
   -La probabilità che riporti il mese di "Luglio" è di 20/240, cioè lo 0.083%.
   -La probabilità che riporti il mese di "Dicembre" 2012 è di 4/240, cioè lo 0.016%.

8) Ho poi creato una nuova colonna contenente i prezzi medi, per farlo ho diviso tutti gli elementi della colonna volume per tutti quelli della colonna sales. Ho moltiplicato per un milione perché gli elementi di volume sono espressi in milioni di dollari.

9) Ho anche creato una colonna che mostri l'efficacia degli annunci di vendita. Per farlo ho calcolato il tasso di conversione dividendo il numero totale di vendite per quello di annunci attivi.

city	media_tasso_conversione
Beaumont	10.6
Bryan-College Station	14.7
Tyler	9.35
Wichita Falls	12.8

Confrontando la media dei tassi di conversione per le quattro città osserviamo che Bryan-College Station è la città con gli annunci più efficaci, mentre Tyler è quella con gli annunci meno efficaci.

10) Ho poi creato due sommari usando dplyr, entrambi mettono a confronto media e deviazione standard. Nel primo i dati sono raggruppati per città e anno:

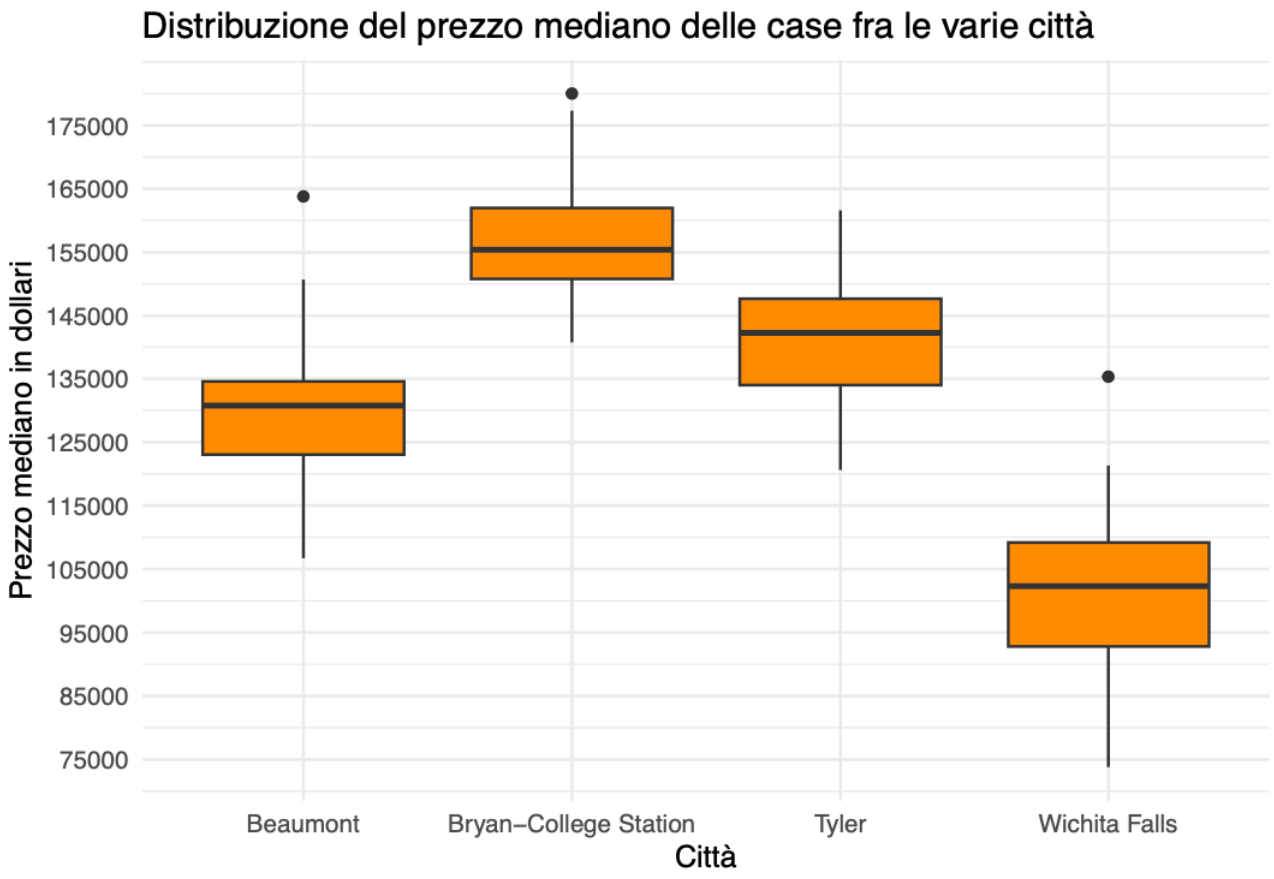
city	year	sales_media	volume_media	listings_media	sales_sd	volume_sd	listings_sd
<chr>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Beaumont	2010	156.1667	22.65383	1731.0833	36.92458	4.954348	101.90768
Beaumont	2011	144.0000	21.09525	1747.9167	22.65552	4.300521	74.67439
Beaumont	2012	171.9167	24.46767	1691.3333	28.38840	4.921853	54.55662
Beaumont	2013	201.1667	30.30917	1639.5833	37.73070	6.436956	54.91888
Beaumont	2014	213.6667	32.13208	1586.6667	36.48993	7.049593	57.37331
Bryan-College Station	2010	167.5833	28.73117	1562.4167	70.75368	10.817885	165.15748
Bryan-College Station	2011	167.4167	28.93083	1606.1667	62.19246	10.313145	156.08962
Bryan-College Station	2012	196.7500	35.35908	1609.5000	74.28217	13.490290	152.98455
Bryan-College Station	2013	237.8333	45.12408	1406.0833	95.84726	19.540786	228.35517
Bryan-College Station	2014	260.2500	52.81283	1106.5000	86.69185	17.970684	127.44446
Tyler	2010	227.5000	36.34658	3051.0833	48.97959	8.394305	226.42577
Tyler	2011	238.8333	38.55333	3069.6667	49.62007	9.405578	184.60195
Tyler	2012	263.5000	44.01042	2910.4167	46.40239	10.229980	122.95192
Tyler	2013	287.4167	50.32492	2823.7500	53.04965	10.325629	159.71971
Tyler	2014	331.5000	59.60167	2670.3333	56.85308	12.761310	172.17028
Wichita Falls	2010	123.4167	14.97200	959.4167	26.61667	4.065984	45.17332
Wichita Falls	2011	106.2500	12.05183	974.8333	19.76280	2.515401	58.58922
Wichita Falls	2012	112.4167	13.23308	896.0000	14.24754	2.662253	41.22003
Wichita Falls	2013	121.2500	14.85142	841.0000	26.00393	3.106485	52.91159
Wichita Falls	2014	117.0000	14.54250	876.6667	21.09287	3.130796	73.05332

Nel secondo per anno e mese:

year	month	sales_media	volume_media	listings_media	sales_sd	volume_sd	listings_sd
<int>	<int>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
2010	1	105.25	15.93775	1616.50	36.60943	6.922791	783.9211
2010	2	121.75	19.22425	1675.75	40.26061	8.535588	779.9484
2010	3	188.75	27.96900	1735.25	43.59950	7.298540	738.8362
2010	4	229.00	33.79900	1902.50	63.95311	13.280337	871.1883
2010	5	232.75	35.92225	1868.25	58.84089	13.236626	947.4173
2010	6	216.50	34.54775	1914.25	71.44462	13.560836	984.9468
2010	7	178.00	26.69100	1942.00	62.50867	12.121784	959.9205
2010	8	184.50	28.58900	1929.50	45.00000	10.666873	954.2175
2010	9	149.50	22.20650	1925.50	47.19816	7.228736	972.0838
2010	10	141.25	21.79100	1867.50	45.70467	8.068802	917.5832
2010	11	125.75	18.41575	1815.50	30.99866	5.684010	879.6384
2010	12	151.00	23.01750	1719.50	40.70217	7.441038	826.7162
2011	1	106.25	15.16450	1741.00	27.03547	5.313391	800.9024
2011	2	117.25	17.34475	1785.25	44.25965	8.107725	833.7063
2011	3	167.00	25.39150	1888.50	52.42773	10.674532	887.1056
2011	4	179.00	27.56675	1948.00	58.64583	11.256460	914.6573
2011	5	195.00	30.50900	1997.50	70.28039	15.375147	922.9047
2011	6	221.25	34.74650	1972.25	93.93038	18.521100	930.5813
2011	7	203.00	32.71375	1944.00	69.95713	14.836718	943.6444
2011	8	196.50	30.20400	1898.00	70.27802	14.510272	940.5683
2011	9	157.25	23.77575	1855.00	67.60855	12.214511	887.7466
2011	10	148.50	22.06850	1808.25	57.57604	10.280971	890.4644
2011	11	137.25	21.57200	1727.50	49.37864	10.695940	833.0564
2011	12	141.25	20.83675	1630.50	48.25194	8.284211	791.3859
2012	1	124.75	17.44775	1700.75	29.78115	6.837480	814.4947
2012	2	143.50	20.62600	1754.50	57.61076	10.198258	823.8529
2012	3	177.75	27.28225	1822.75	66.69020	12.719385	811.8002
2012	4	186.75	28.03200	1854.00	52.77863	10.981393	834.8185
2012	5	220.50	36.88150	1863.25	90.71751	19.394133	849.5153
2012	6	224.50	37.33375	1858.25	86.18004	18.633920	873.0400
2012	7	232.00	38.11575	1857.75	89.81462	20.164871	887.3719
2012	8	238.50	38.56775	1794.00	87.99432	19.762346	892.0617
2012	9	176.75	28.18975	1771.50	78.20646	15.251248	852.7878
2012	10	185.50	28.75100	1733.25	79.80601	15.203367	839.9684
2012	11	162.50	25.63675	1700.25	37.24245	7.173133	821.1977
2012	12	160.75	24.34650	1611.50	52.20073	10.330214	759.5668
2013	1	144.00	22.93475	1644.75	49.22059	9.591138	748.6046
2013	2	148.25	22.48650	1683.75	54.90219	11.426154	746.3692
2013	3	203.50	31.37550	1743.25	64.04426	13.893366	793.4471
2013	4	219.50	36.40525	1797.75	74.54082	18.420387	836.3852
2013	5	264.25	46.11850	1771.50	90.36362	22.272422	853.9924
2013	6	261.25	46.07875	1761.50	108.21699	24.932241	875.5313
2013	7	281.75	47.13775	1733.75	122.70391	26.927092	915.1089
2013	8	276.75	46.40925	1710.75	92.01585	20.868560	899.3140
2013	9	203.50	34.19525	1652.75	66.00253	14.307844	914.7066
2013	10	184.50	31.14725	1614.50	65.97727	15.597292	897.0561
2013	11	172.50	27.54875	1558.50	65.07688	12.784377	837.1063
2013	12	183.25	29.99125	1458.50	64.04881	12.059098	766.7170
2014	1	156.75	23.51900	1532.25	61.34805	12.074357	793.8226
2014	2	173.50	28.57600	1563.25	62.21736	13.162092	790.5101
2014	3	210.25	34.90525	1593.75	85.35563	17.803333	814.8412
2014	4	244.25	40.71925	1626.25	84.10063	20.136292	827.1817
2014	5	281.75	49.07925	1618.75	112.15577	26.326048	806.1585
2014	6	294.25	53.80900	1660.00	134.62386	30.670720	851.4368
2014	7	284.00	50.95125	1628.50	122.29745	29.507853	889.2669
2014	8	261.00	46.30075	1599.25	89.70693	22.953933	847.1428
2014	9	224.75	39.62850	1539.75	103.54186	23.355544	809.5638
2014	10	239.75	41.63525	1528.25	106.31518	21.281137	772.6195
2014	11	186.25	30.86250	1461.75	84.50000	17.263882	728.5403
2014	12	210.75	37.28125	1368.75	91.73649	19.756554	675.7817

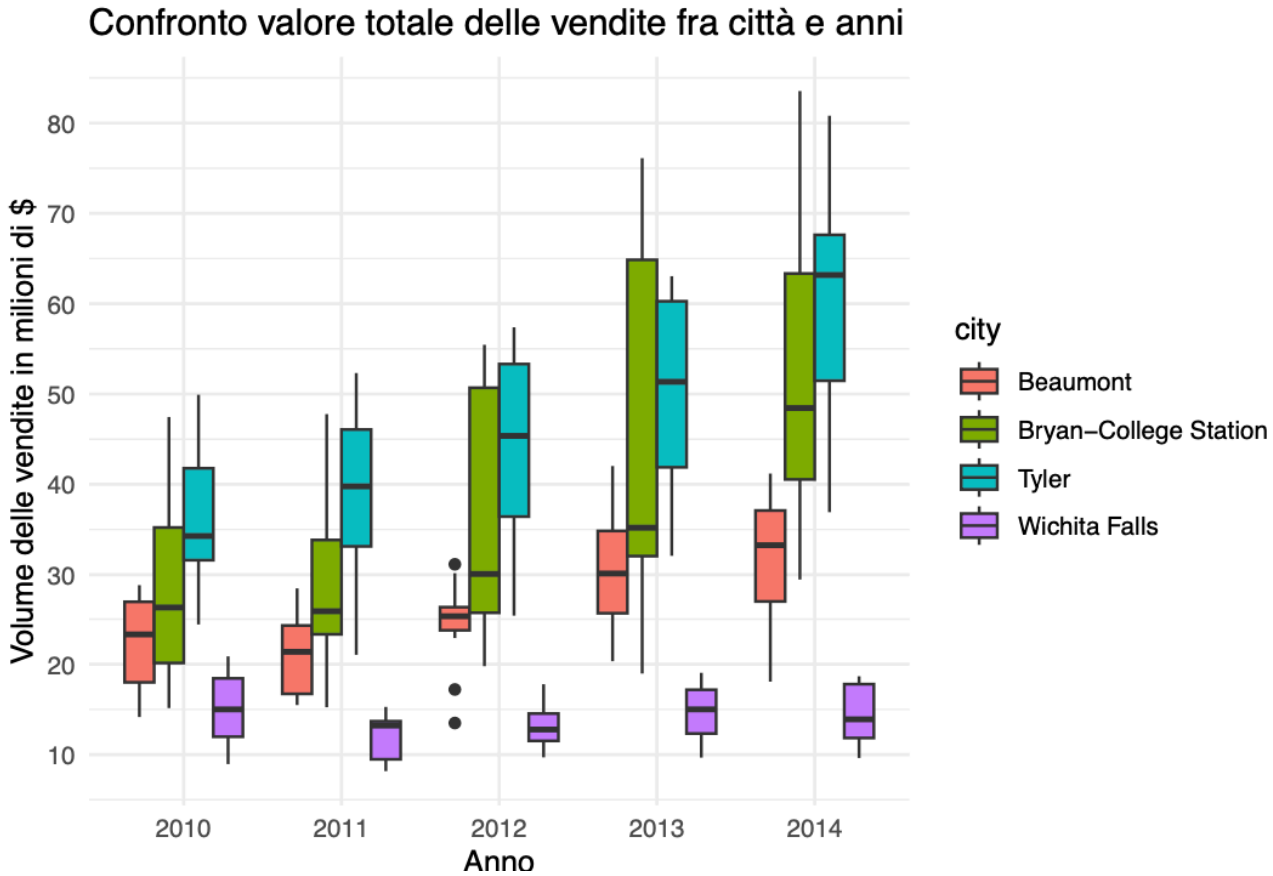
Avvalendomi di ggplot2 sono andato a realizzare dei grafici per visualizzare e comprendere al meglio i dati trattati.

1) Utilizziamo i boxplot per confrontare la distribuzione del prezzo mediano delle case tra le varie città.



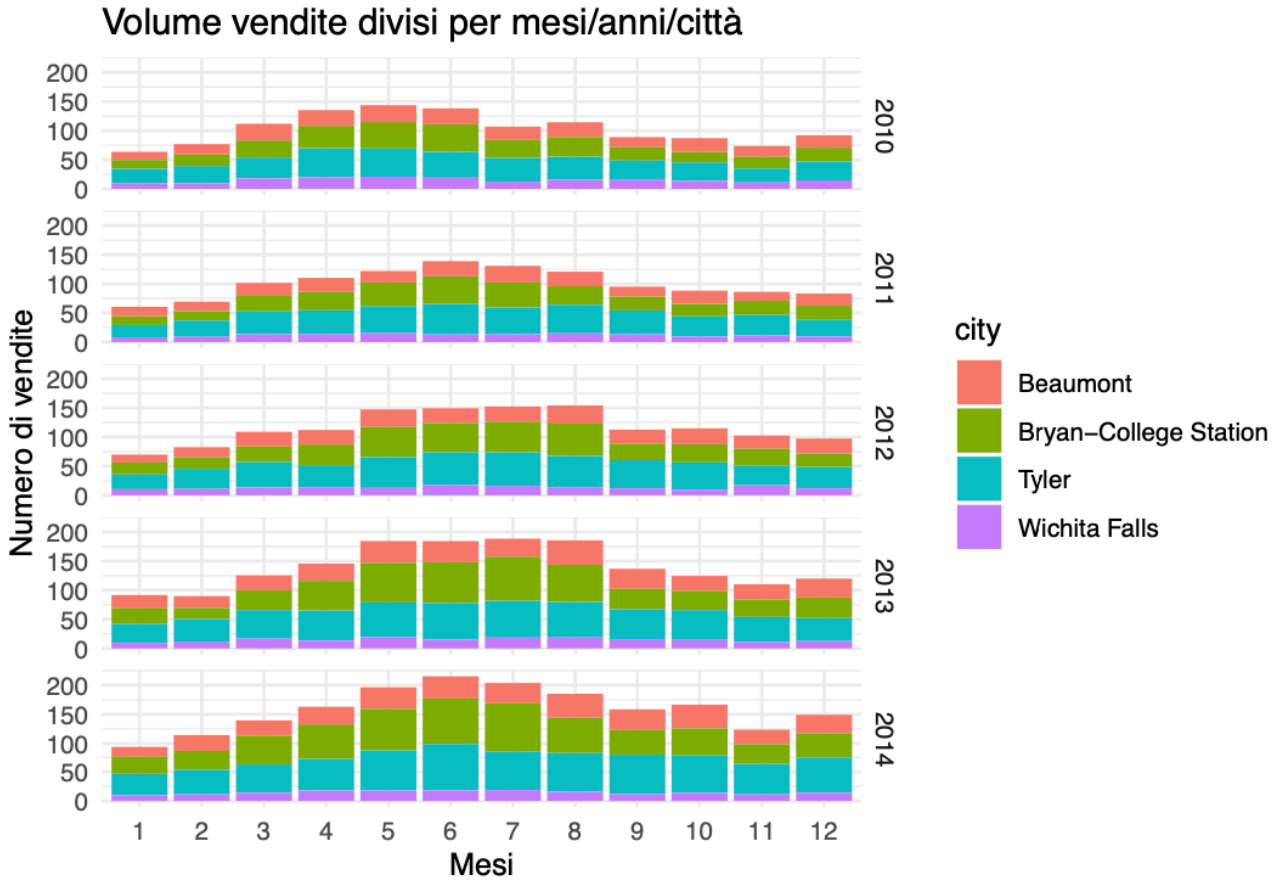
Dal grafico notiamo che Wichita Falls è la città con i prezzi più variabili, ha però un prezzo mediano di acquisto inferiore rispetto alle altre città. Al contrario, Bryan-College Station ha il prezzo mediano d'acquisto più alto presentando però prezzi meno variabili rispetto alle altre città.

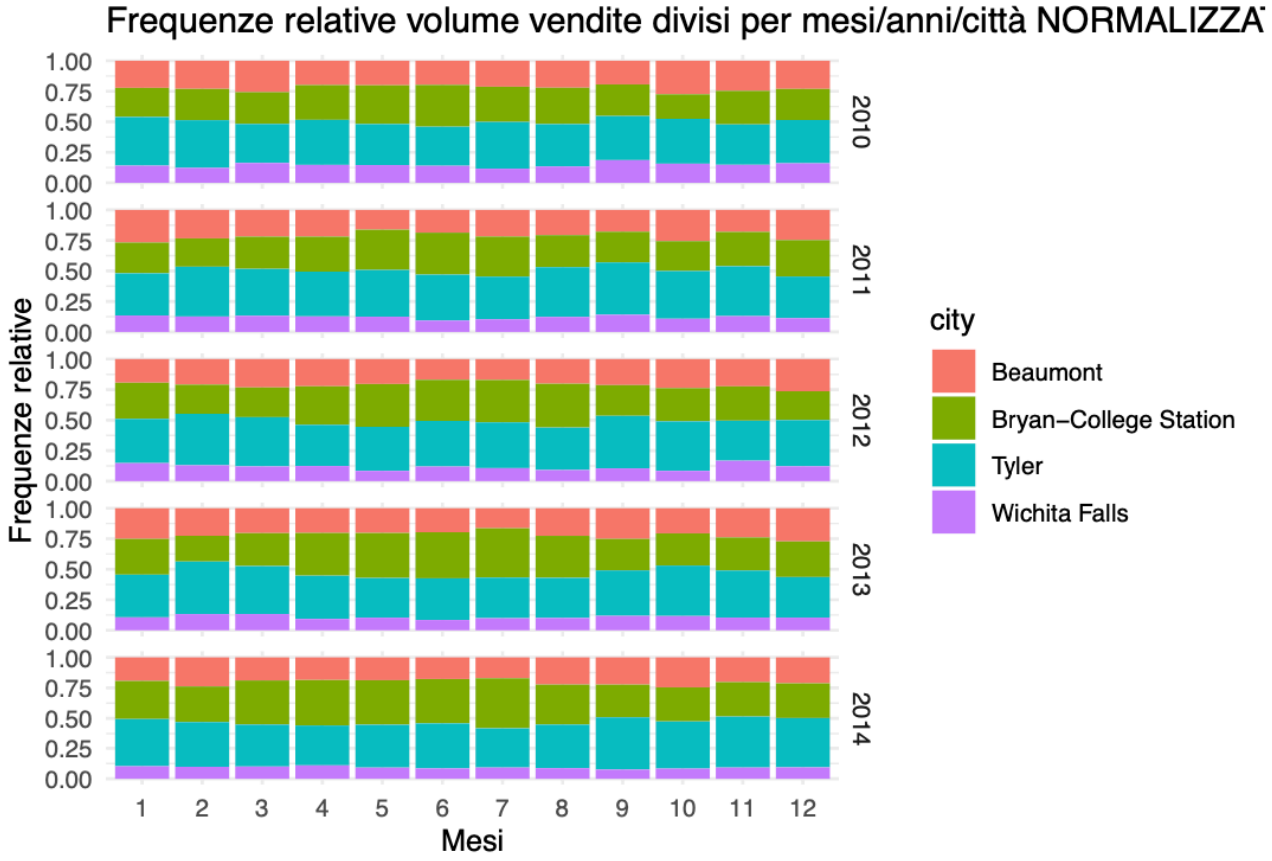
2) Adesso utilizzeremo i boxplot per confrontare la distribuzione del valore totale delle vendite tra le varie città ma anche tra i vari anni.



Dal grafico notiamo che ad eccezione di Wichita Falls i volumi sono aumentati nel tempo, complice anche l'aumento generale dei prezzi delle case. Tyler è stata la città più redditizia, Wichita Falls quella con i volumi più costanti e Bryan-College Station quella con i volumi più variabili.

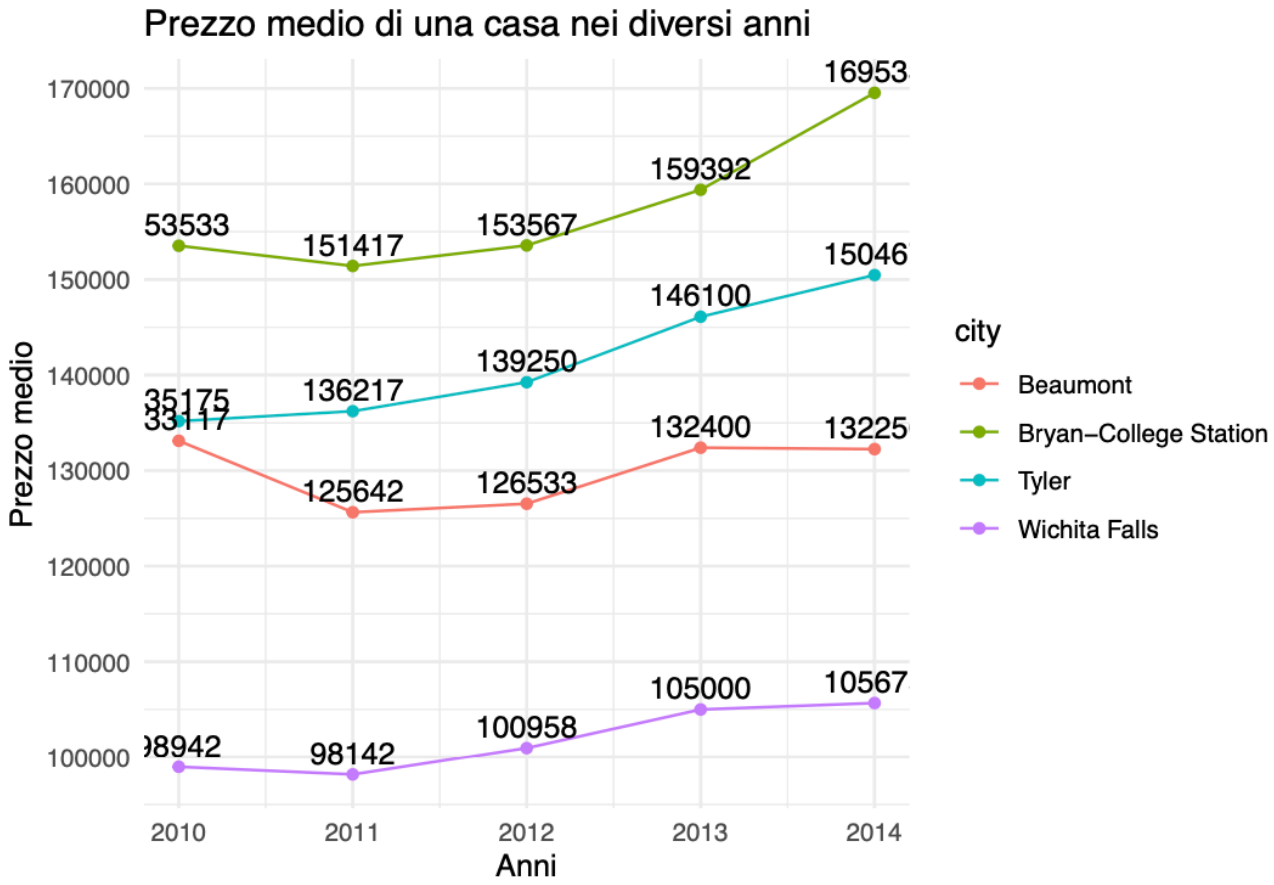
3) Avvaliamoci adesso di un grafico a barre sovrapposte per per confrontare il totale delle vendite nei vari mesi e anni, sempre considerando le città.





Entrambi i grafici ci sono molto utili, dal primo vediamo che si hanno più vendite nei mesi primaverili/estivi rispetto a quelli autunnali/invernali. Dal secondo invece ci rendiamo conto che la città in cui si vendono più case è Tyler, seguita da Bryan-College Station, Beaumont ed infine Wichita Falls.

4) Come ultimo grafico andremo a realizzare un line chart raffigurante il prezzo di vendita di una casa nelle varie città nel corso degli anni.



-Wichita Falls si dimostra la città con i prezzi più bassi e stabili nel corso del tempo.
-A Beaumont abbiamo invece assistito prima ad una discesa dei prezzi durata due anni per poi risalire e stabilizzarsi.
-Tyler ha avuto un costante innalzamento dei prezzi.
-Bryan-College Station si rivela la città con i prezzi più alti con un'impennata fra il 2013 e il 2014.