PROGETTO STATISTICA DESCRITTIVA

1. Importo i dati e carico le librerie necessarie
2. Indico il tipo di variabili contenute:

city = qualitativa nominale

year = quantitativa continua (su scala di intervalli)

month = qualitativa ma espressa in numeri

sales = quantitativa discreta (su scala di rapporti)

volume = quantitativa continua (su scala di rapporti)

median= quantitativa continua anche se assume solo valori interi (su scala di rapporti)

listing = quantitativa discreta (su scala di rapporti)

month inventory = quantitativa continua (su scala di rapporti)

1. Applico la funzione *table* alle variabili qualitatitive (city, *year*, *month)* ed è risultato che hanno la stessa frequenza per tutte le modalità, quindi si tratta di variabili senza moda.

Poiché non ha senso calcolare la mediana per le variabili *city* e *month*, creo un nuovo data frame che non le contiene. Applico un ciclo for per il calcolo della mediana.

Un altro modo è utilizzare la funzione *quantile* ed estrarre il 50 esimo percentile. Infine ho calcolato i decili.

La funzione *summary* restituisce gli indici di posizione. È utile per il calcolo della media. Applico questa funzione alle variabili *sales, volume, median\_price\_listings* e *month\_inventory* in quanto non ha senso calcolare la media per le variabili qualitative per le quali ho precedentemente creato la distribuzione di frequenza mediante la funzione *table*.

Passando alle misure di variabilità, creo un nuovo data frame ed escludo le variabili year, city e month.

Uso un ciclo for per calcolare il range di variazione, il range interquartile, la varianza e la deviazione standard.

Infine calcolo l’indice di simmetria con skewness.

Per i confronti fra variabili si veda il punto 4.

1. La variabile più asimmetrica è *volume* (risultato della funzione skewness).

Per un confronto sulla variabilità, calcolo il coefficiente di variazione: *volume* è la variabile con variabilità più elevata.

1. Scelgo la variabile *Sales* che divido in 5 classi. Dal barplot noto che le barre non sono tutte uguali (equidistribuzione) e neanche i valori sono concentrati in un’unica barra (massima concentrazione). Posso quindi escludere i due casi limite. Calcolo l’indice di Gini che restituisce 0.91 che è un valore alto di eterogeneità essendo il massimo pari ad 1.
2. Nel caso di *city*, l’indice di Gini è pari ad 1. Mi aspettavo questo risultato in quando avevo già notato l’equidistribuzione di questa variabile, come anche per *month* e *year*.
3. La probabilità si calcola:

casi favorevoli/ casi possibili

Utilizzando la funzione table:

* La probabilità che presa una riga a caso la città sia “Beaumont” è 60/240 , quindi 25%
* La probabilità che presa una riga a caso il mese sia “Luglio” è 20/240, quindi 8,3%
* La probabilità che presa una riga a caso il mese sia “Dicembre 2012” è 4/240, quindi 1,66%

1. Creo una nuova colonna nel data frame my\_data che chiamo *mean\_price* che è pari a *volume/sales*.

La nuova variabili indica il prezzo medio. Ho moltiplicato per 1 milione per poter fare il confronto con *median price*. I valori ovviamente sono diversi, ma simili perché entrambi sono due indici di posizione che assumo valori “al centro” della distribuzione. (Da tener presente che la mediana è un indice più robusto della media).

1. Ho creato una variabile chiamata *efficacia* pari a *listings/sales*. Ad esempio per la prima osservazione, si registrano 83 vendite e un numero totale di annunci attivi pari a 1533.

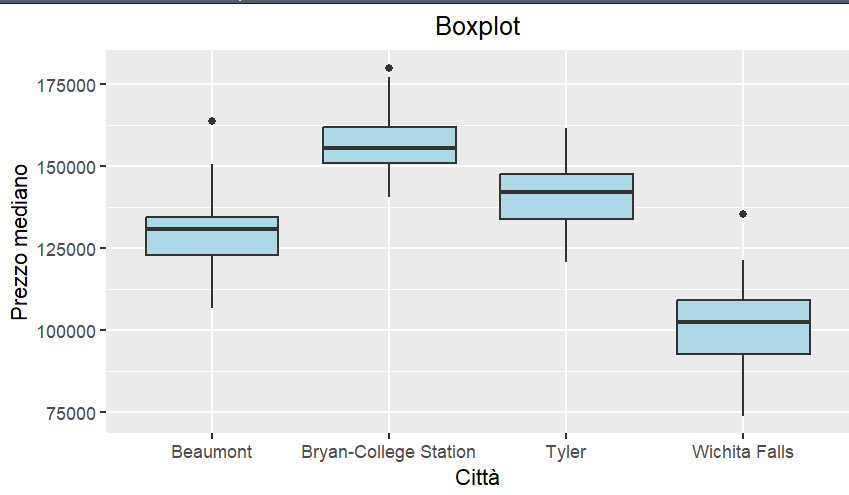
1533/83 è 18,5. Ciò vuol dire che sono necessari quasi 19 annunci per poter vendere un immobile.

1. Scelgo come variabili *sales* e *volume*. Sfruttando la notazione suggerita, calcolo media e deviazione standard condizionatamente alla città, agli anni e ai mesi. Da questi risultati, risulta che:

* le città che mediamente hanno i valori più alti di entrambe le variabili sono in ordine Tyler, Bryan-College Station, Beaumont e Wichita Falls (anche i valori di sd sono in questo ordine);
* negli anni c’è stato un aumento di entrambe le variabili (sono aumentati anche i valori di sd)
* entrambe le variabili assumono valori maggiori nella stagione primaverile-estiva (stesso andamento per sd).

GRAFICI

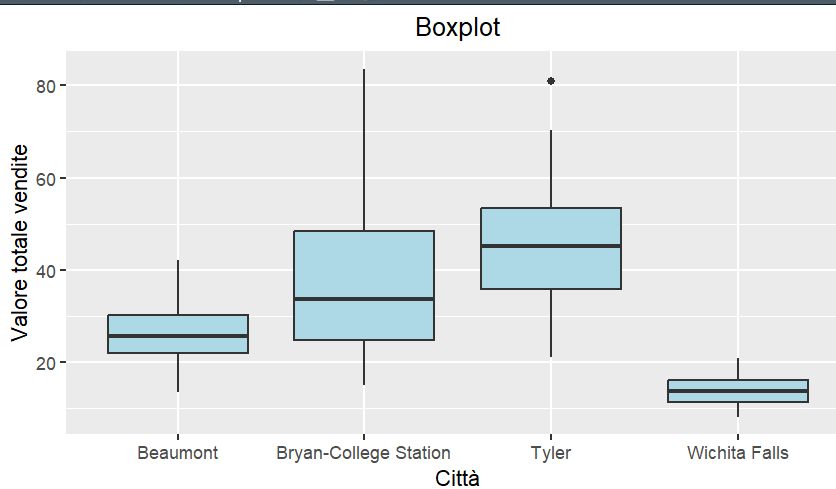
1. Creo 4 boxplot per confrontare la distribuzione del prezzo mediano delle case tra le varie città.



* Per questa variabile, Bryan-College Station è la città con media più alta seguita da a Tyler, Beaumont e Wichita Falls.
* Tre di questi boxplot presentano valori anomali.
* Nessuna variabile è perfettamente simmetrica. La funzione skewness permette di calcolare l’asimmetria. Tyler è la più simmetrica, Bryan-College Station la più asimmetrica. È un risultato che mi aspettavo dato che Tyler non presenta valori anomali a differenza delle altre variabili che hanno valori anomali su una coda.

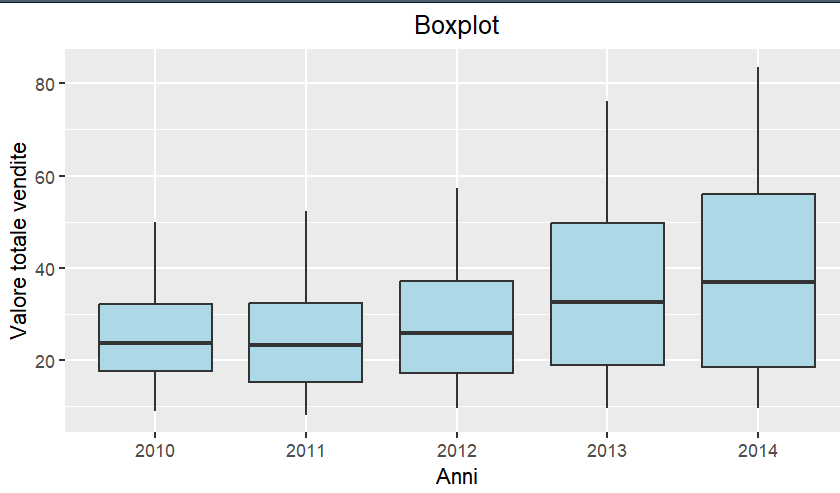
Non è molto evidente capire quale boxplot presenti la minore variabilità, mi aiuto allora con il calcolo della deviazione standard tramite la funzione *tapply* e risultata che la variabile con minore variabilità è Bryan-College Station.

1. Creo dei boxplot per confrontare il valore totale delle vendite rispetto alle città.



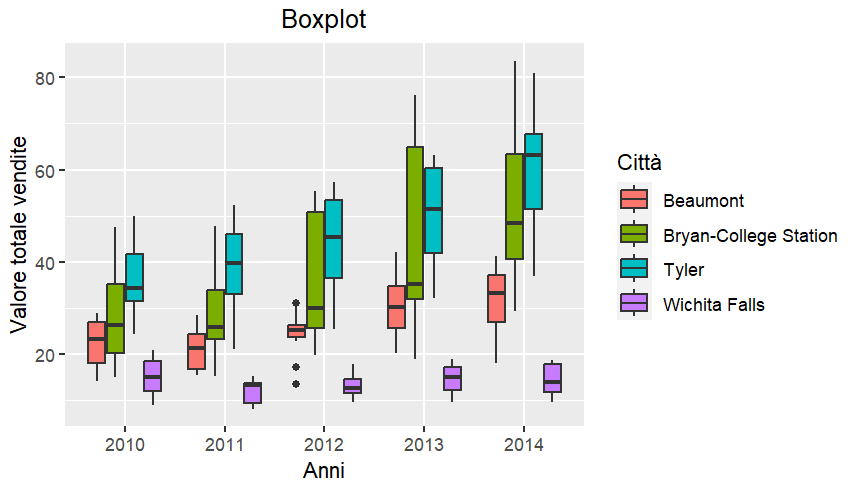
I boxplot ottenuti chiariscono i risultati ottenuti nel punto 10 poiché avevo scelto la variabile *Volume*. In questo caso  è più semplice capire quale boxplot presenti la maggiore variabilità, asimmetria e intervallo interquartile: si tratta di Bryan-College Station. Tale risultato è confermato dal calcolo delle corrispondenti funzioni.

Poi, creo altri boxplot per confrontare il valore totale delle vendite rispetto agli anni.



Anche in questo caso i boxplot chiariscono i risultati ottenuti al punto 10. È evidente che con il passare degli anni è aumentata la media, la variabilità e l'intervallo di variazione. La distribuzione più simmetrica quella del 2014.

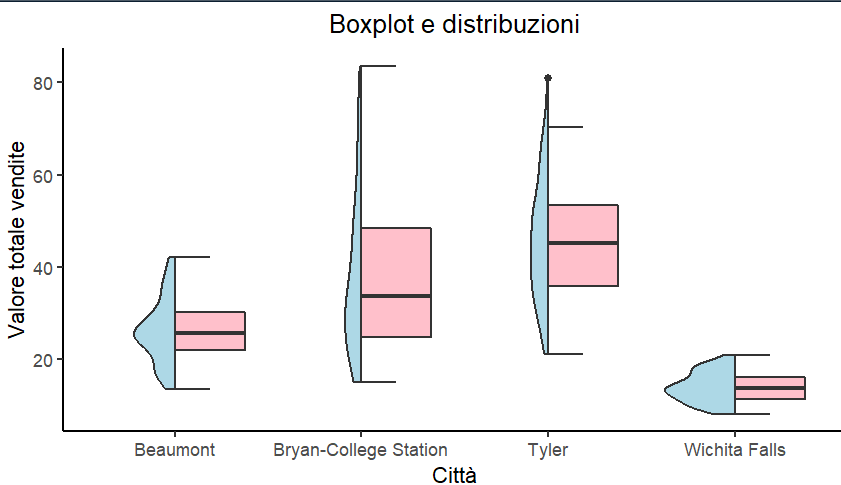
Infine calcolo i boxplot considerando insieme gli anni e le città

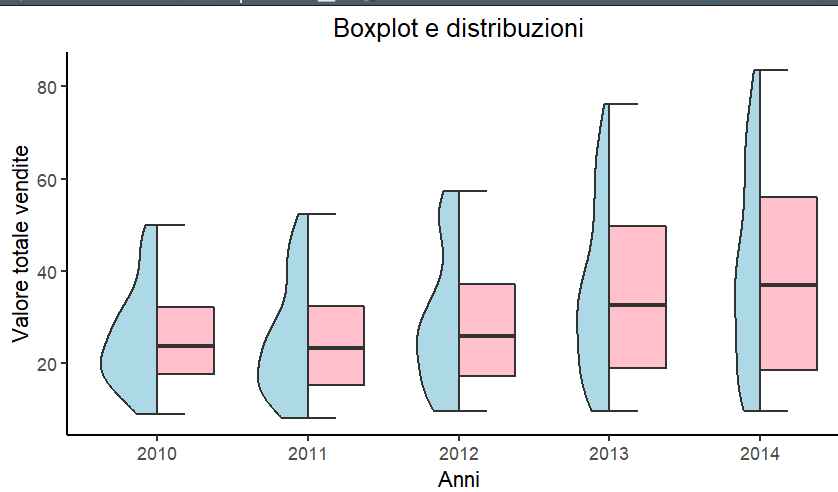


I risultati sono i seguenti:

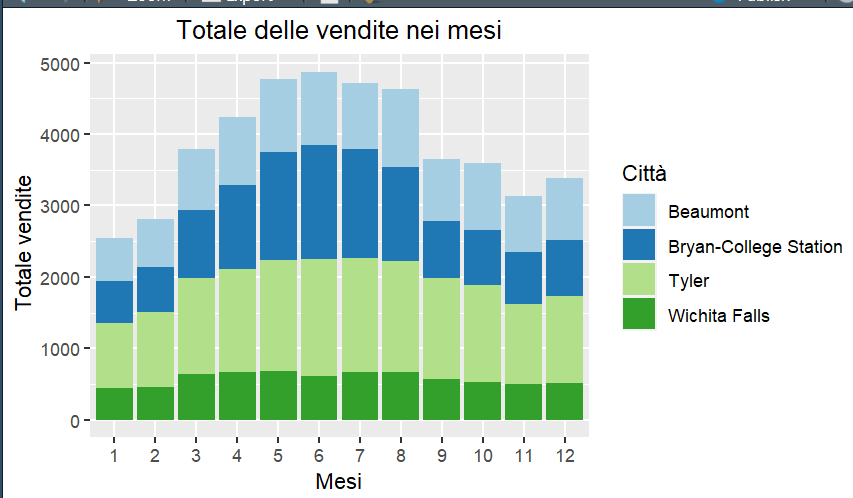
* Con il passare degli anni non sono cambiate le differenze fra le città rispetto alla variabile *volume*: Witchita Falls rimane la città più economica, seguita da Beaumont e Bryan-College Station. Tyler è la più costosa
* La città Bryan\_College Station è la più variabile ogni anno
* La città con maggiore IQR è Bryan-College Station (tranne nel 2011)
* La distribuzione più asimmetrica è quella del 2012 nella città di Beaumont

Per completare l’analisi ho aggiunto distribuzioni affianco ai boxplot per avere una maggiore visibilità sull’asimmetria.





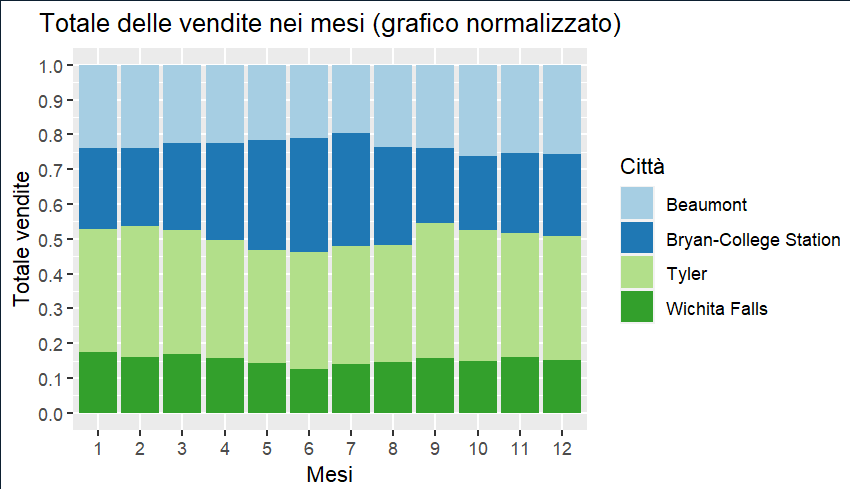
1. Creo un grafico a barre sovrapposte per confrontare la variabile *volume* nei vari mesi e città.



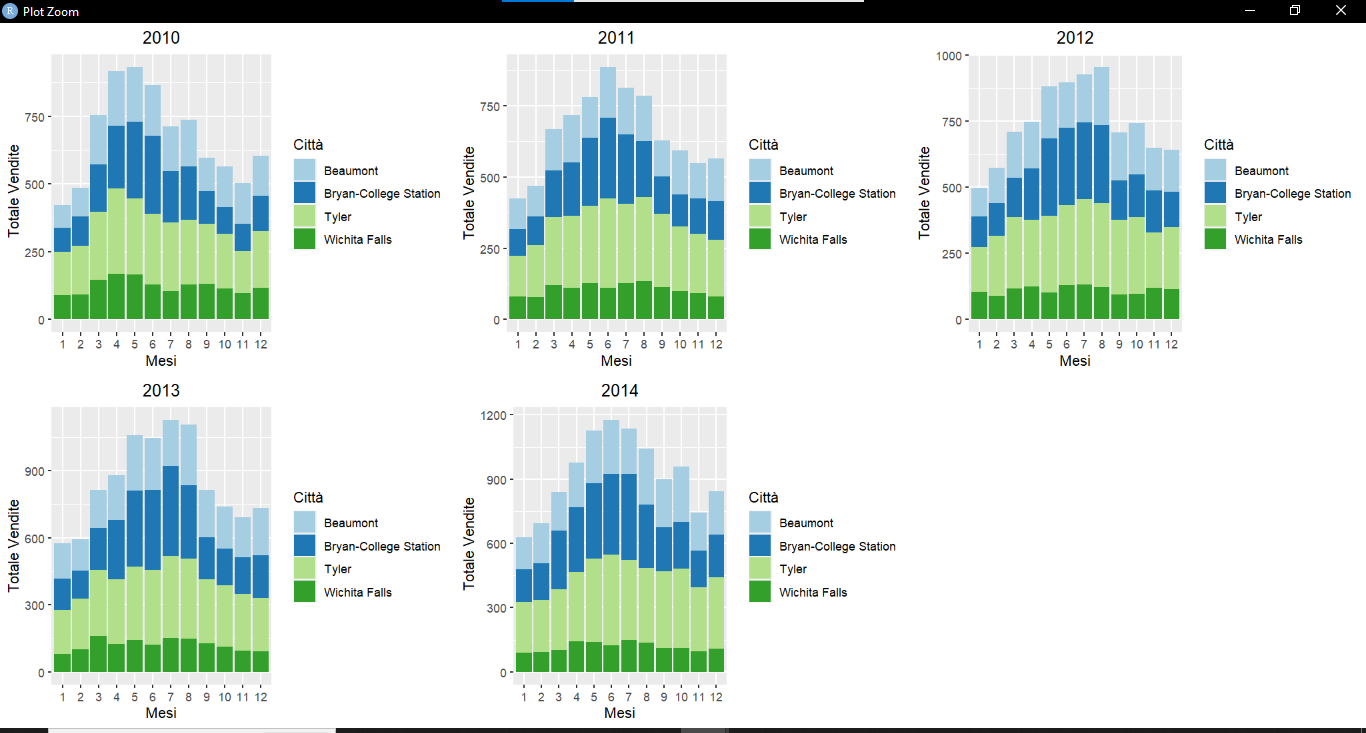
I risultati sono i seguenti:

* Nei mesi primaverili ed estivi c'è un aumento delle vendite (fenomeno stagionale).
* La città con più vendite è Tyler; Wichita con meno vendite nei mesi
* Non si riescono ad apprezzare le differenze anno per anno in quanto si considerano 5 mesi di gennaio, 5 mesi di febbraio…

Anche dal grafico normalizzato si possono comprendere gli ultimi due risultati

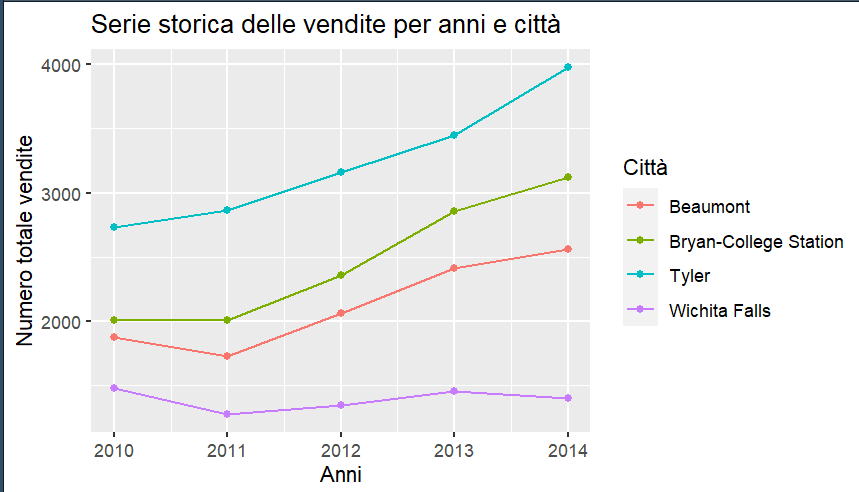


Per poter apprezzare le differenze anno per anno, ho creato 5 grafici a barre non normalizzati e 5 grafici a barre normalizzati. Ho quindi deciso di considerare la variabile *Year* attraverso la creazione di questi altri grafici in modo da creare grafici di facile lettura e che mi permettessero di visualizzare i risultati annuali considerando insieme le variabili *Sales*, *Month*, *City* e *Year*.



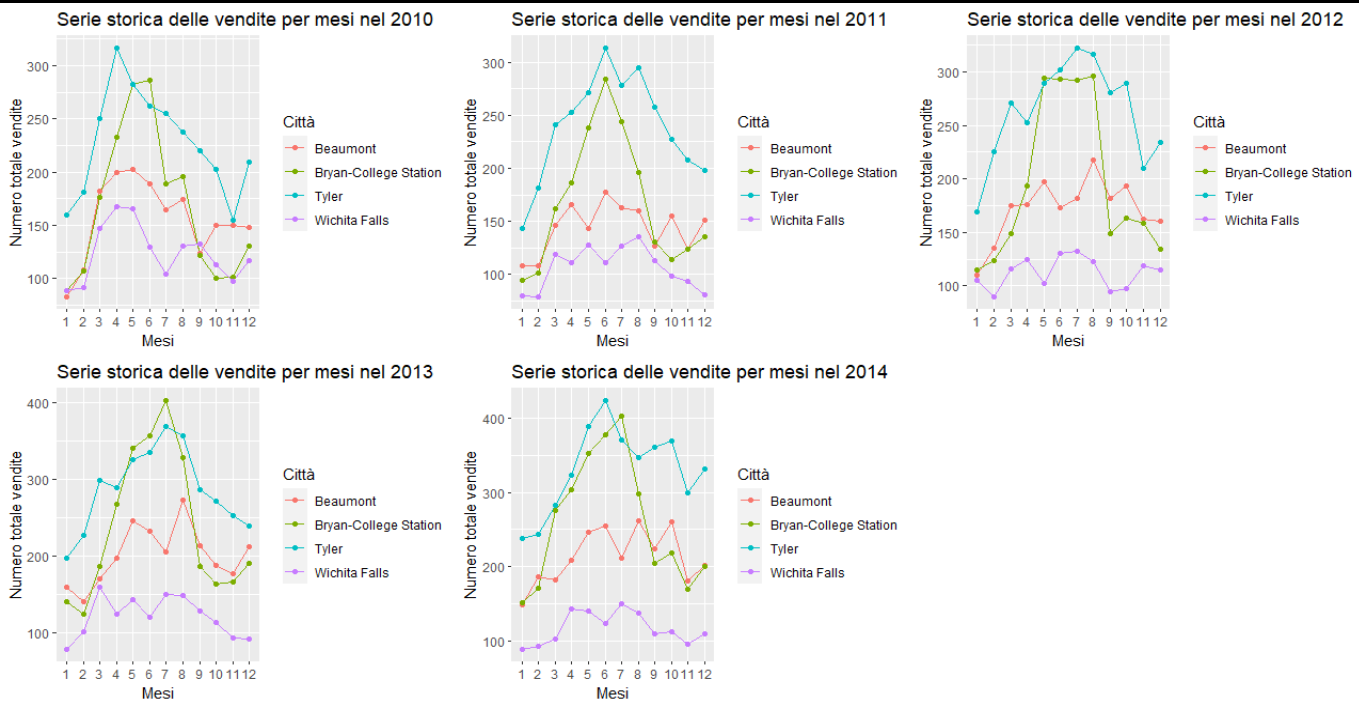
I risultati precedenti ottenuti considerando i 5 anni sono confermati anche ogni anno. Infatti, esiste un fenomeno stagionale che si ripete tutti gli anni con Tyler che rimane sempre la città con il maggiore numero di vendite e Witchita Falls con il numero minore di vendite, mentre Bryan -College Station occupa il secondo posto per la gran parte dei mesi.

1. Creo un data frame che chiamo *tabella* e che mi permette di disporre della somma di *Sales* ogni anno per ogni città. Ho quindi sommato i valori di *Sales* per tutti i mesi di un anno. Questo data frame mi permette di creare un line chart che mostra la serie storia delle vendite per anni e città.



Si evince ancora che il numero totale delle vendite è cresciuto nel tempo. Una piccola eccezione per Wichita Falls per la quale c’è una piccola diminuzione del 2014. Tyler è la città con valori più alti ogni anno. Le linee non si incontrano a dimostrare che ogni anno è rispettato il seguente ordine fra le città per numero totale di vendite: Tyler, Bryan-College Station, Beaumont e Wichita Falls.

Per considerare anche la variabile *month* ho creato 5 line chart ognuno per ogni anno



Si evince che la variabile *Sales* assume valori maggiori nei mesi primaverili e estivi. In qualche mese le linee si incontrano, capita talvolta anche che i valori di Bryan-College Station siano superiori a Tyler, ma se consideriamo l’intero anno Tyler assume in genere i valori più alti.