Progetto – Statistica Descrittiva

Il seguente documento conterrà il progetto di Statistica Descrittiva del master professionale in Data Science di Profession AI.

Task 1 – Il dataset

Il dataset riguarda la vendita di immobili in Texas.
Il dataset è composto da otto variabili:

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

L'import dei dati su R è stato effettuato tramite la seguente linea di codice:
data_texas <- read.csv("realestate_texas.csv")
che ha permesso di leggere l'intero dataset e inserirlo all'interno dell'ambiente di R.

Task 2 – Descrizione del dataset

Il dataset viene esplorato tramite la funzione base di R "head" che ci permette di visualizzare le righe e colonne del dataset. La linea di codice utilizzata è stata:

head(data_texas, 5)

Di seguito un'analisi di ciascuna variabile del dataset:

- city: variabile qualitativa su scala nominale
- year: variabile qualitativa ordinata
- month: variabile qualitativa ordinata
- sales: variabile quantitativa discreta
- volume: variabile quantitativa continua
- median_price: variabile quantitativa continua
- listings: variabile quantitativa discreta
- months_inventory: variabile quantitativa continua

In definitiva, sono presenti 5 variabili quantitative discrete, 2 variabili quantitative continue e 1 variabile qualitativa su scala nominale.

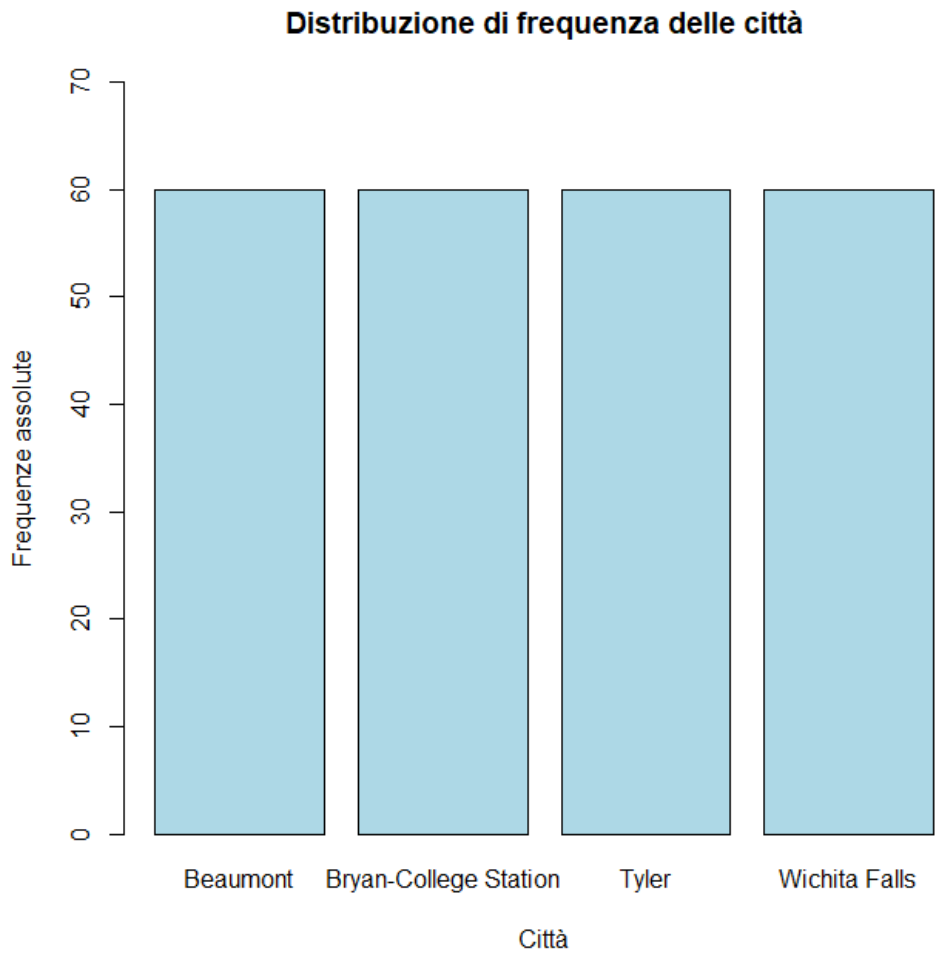
Task 3 – Calcolo degli indici per le variabili

Le variabili individuati per le quali non ha senso procedere al calcolo degli indici di posizione, variabilità e forma sono:

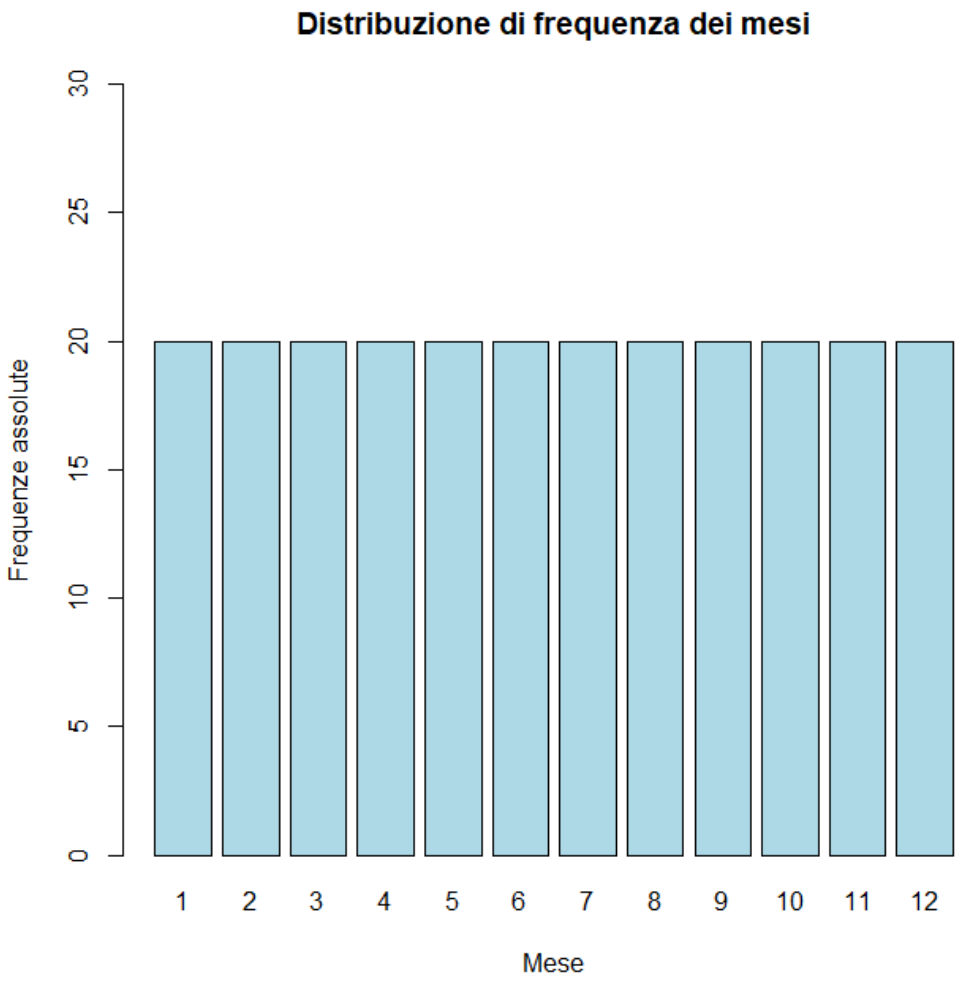
- city: variabile qualitativa su scala nominale
- year: variabile quantitativa discreta
- month: variabile quantitativa discreta

Questo perché non danno informazioni aggiuntive studiandone gli indici essendo variabili che hanno già di loro un ordine naturale. Inoltre, andando a calcolare una media, ad esempio, non risulterebbe corretto poiché nonostante anni e mesi sono variabili numeriche, non possono essere sommate e quindi, sintetizzate.
Per queste variabili sono state calcolate le distribuzioni di frequenza e visualizzate attraverso un grafico di tipo barplot del pacchetto base di R.

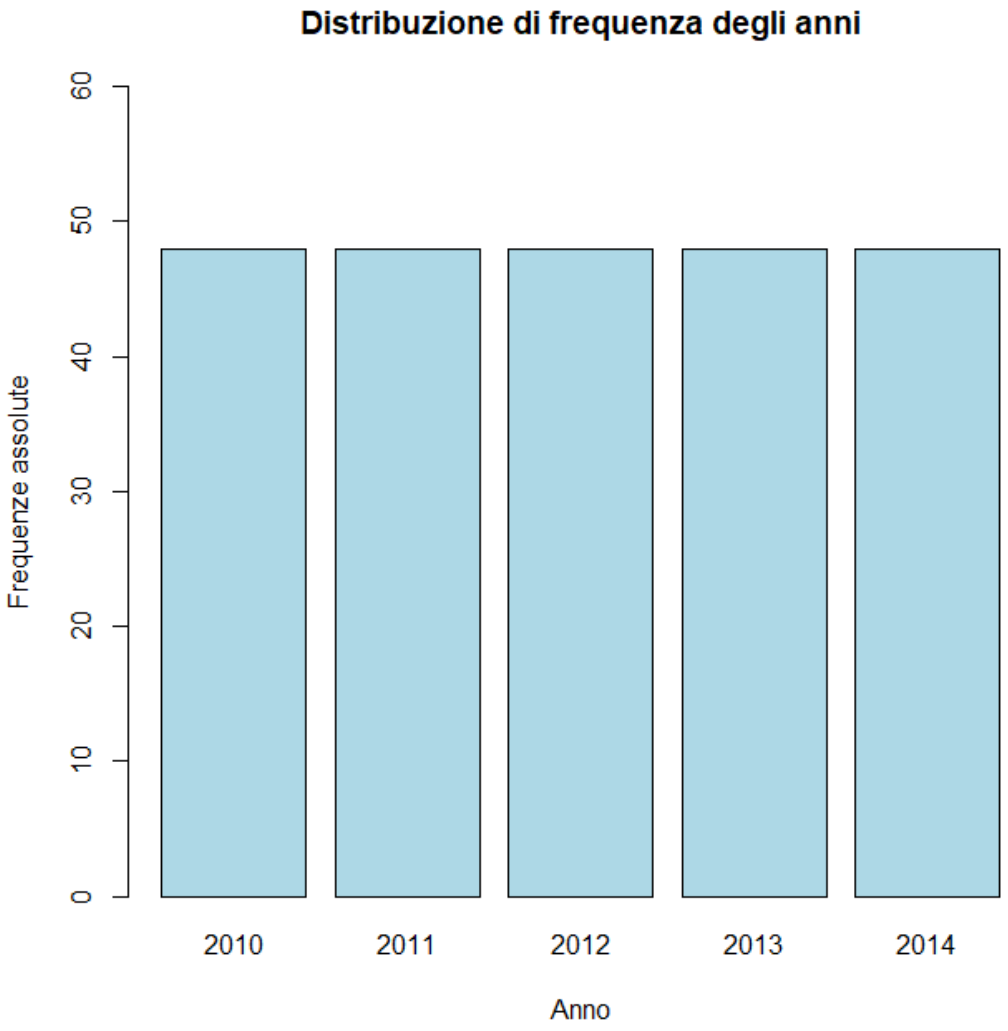
Distribuzione di frequenza per la variabile city



Distribuzione di frequenza per la variabile month



Distribuzione di frequenza per la variabile year



Le variabili utili al calcolo degli indici di posizione, variabilità e forma sono:

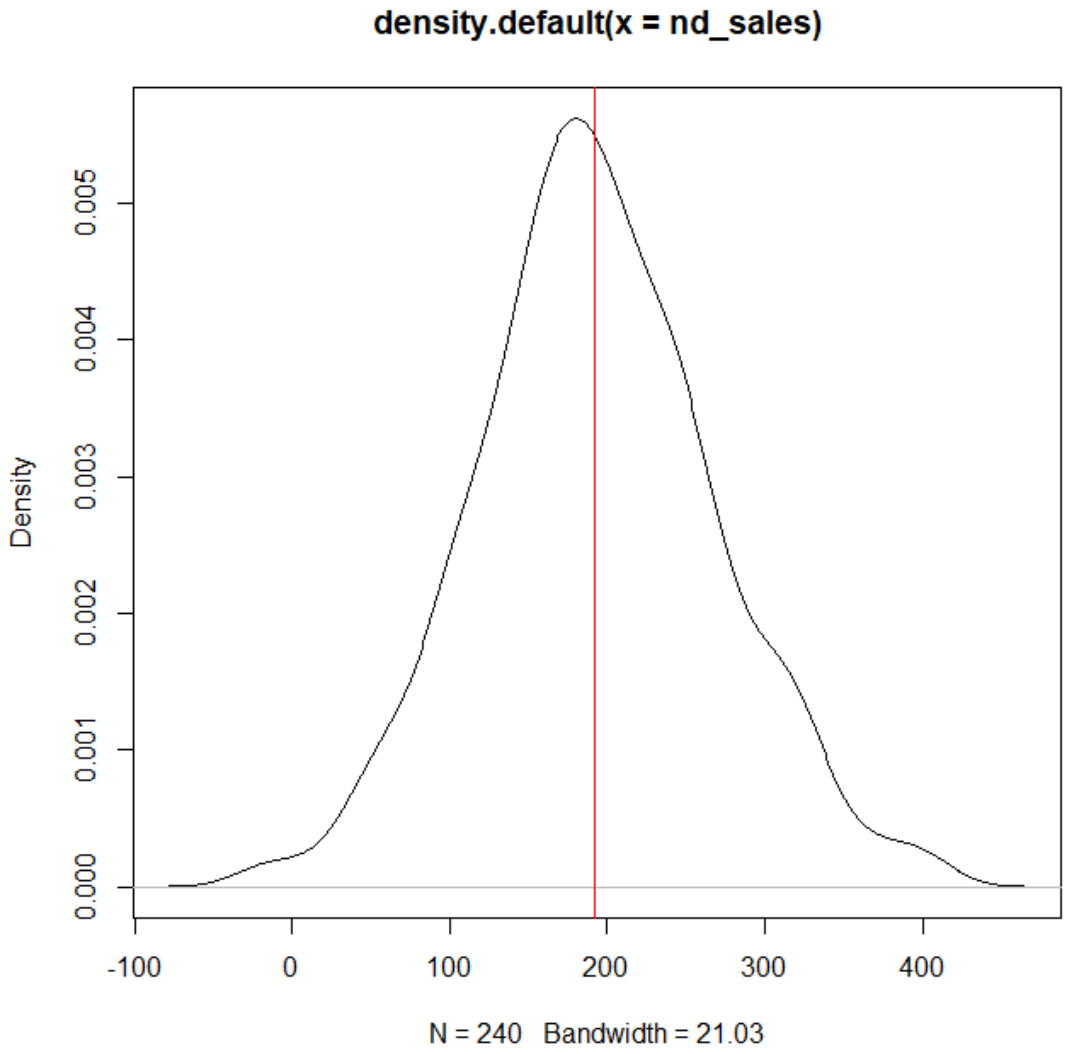
- sales: numero totale di vendite
- volume: valore totale delle vendite in milioni di dollari
- median_price: prezzo mediano di vendita in dollari
- listings: numero totale di annunci attivi
- months_inventory: quantità di tempo necessaria per vendere tutte le inserzioni correnti al ritmo attuale delle vendite, espresso in mesi

Di queste sono stati calcolati gli indici di posizione singolarmente attraverso le funzioni base di R, gli indici di variabilità attraverso delle funzioni implementate custom ed infine, gli indici di forma costruendo la distribuzione normale delle variabili.
Inoltre, la distribuzione normale è stata plottata come grafico di densità di probabilità.

Tabella riassuntiva degli indici di posizione, variabilità e forma per le variabili: sales, volume, median_price, listings e month_inventory

	sales	volume	median_price	listings	month_inventory
Min	79	8.166	73800	743	3.4
Max	423	83.547	180000	3296	14.9
Mean	192.3	31.00519	132665.4	1738.021	9.1925
Median	175.5	27.062	134500	1618.5	8.95
Range	344	75.381	106200	2553	11.5
Variance	6317.865	276.1154	511433095.65972	564208.3	5.28477
Standard deviation	79.485	16.61672	22614.88659	43	2.29886
Coefficient variance	41.42203	53.70536	17.08218	43.30833	25.06031
Fisher index	0.71362	0.87921	0.87921	0.64544	0.04071
Kurtosis index	-0.33552	0.15056	0.64272	-0.81015	-0.19794

Grafico di densità della variabile sales



Task 4 – Verifica quale variabile risulta avere una variabilità maggiore

L'analisi sulla verifica della variabile con variabilità maggiore è stata quella di creare i boxplot delle singole variabili considerate valutabili dal punto di vista di calcolo degli indici, per confrontare le distribuzioni. Successivamente, si è andato a confrontare la deviazione standard di ciascuna variabile ed è emerso che in particolare, "listings" e "median_price" presentano una deviazione standard in ordini di grandezza molto più grandi rispetto alle variabili "months_inventory", "sales" e "volume". Per questo motivo, si è ritenuto necessario procedere al confronto dei coefficienti di variazione che permettono di valutare la variabilità in maniera relativa tra diverse distribuzioni. In questo modo, è stato possibile confrontarne la variabilità ed è emerso che la variabile con variabilità maggiore è: "volume", la quale presenta un coefficiente di variazione di 53.7, cioè una deviazione standard che risulta essere il 57% della media.

Per quanto riguarda invece della variabile più asimmetrica è stato effettuato un confronto degli indici di Fisher tra le varie variabili e si è andato ad individuate quello con un valore maggiore rispetto a tutti gli altri.

Dal confronto è emerso che la variabile con asimmetria maggiore è la: "volume", la quale presenta un indice pari a 0.88. Il valore dell'indice va ad indicare che la distribuzione della variabile "volume" risulta essere

asimmetrica positiva. Inoltre, verificando anche l'indice di Curtosi, la distribuzione risulta essere anche leptocurtica poiché ha un indice pari a 0.15 (code sottili e nella zona centrale curva appuntita e alta).

Task 5 – Variabile quantitativa divisa in classi

La variabile scelta per la suddivisione in classi è stata "sales".

La scelta delle classi di 0 – 500 con passo da 100 è perché considerato il minimo della distribuzione di 79 e il massimo di 429, non avrebbe avuto senso creare delle classi da 0 a 500 con un passo da 50, si avrebbe ottenuto la prima classe 0 – 50 completamente vuota, per questo motivo, si è optato per passare da 0 – 500 con passo 100.

Al dataset principale si è aggiunta una nuova colonna che inserisse tutte le classi della variabile sales in modo tale da calcolarne le frequenze successivamente.

Una volta stabilite le classi e aggiunte al dataset, sono state calcolate le frequenze assolute, relative, cumulate e relative cumulate per le classi costruite.
Tutte le frequenze sono state inserite all'interno di un dataframe con cui si sono potuti realizzare i grafici a barre che visualizzano sia le frequenze assolute che quelle relative suddivise per classi

Infine, si è calcolato l'indice di Gini per la variabile Sales, il quale risulta essere pari a 0.9.
Il suo valore ci suggerisce che la distribuzione della variabile continua sales sembra essere quasi equidistribuita lungo la sua distribuzione.

Task 6 – Indice di Gini per la variabile city

Considerando che per definizione l'indice di eterogeneità di Gini misura la propensione di una variabile qualitativa ad assumere le sue diverse modalità, per la variabili city si è in presenta di una equidistribuzione, cioè l'indice di Gini è esattamente uguale ad 1.

Questa deduzione nasce nel considerare la frequenza assoluta della variabile city, la quale risulta essere pari a 60 per ciascuna modalità (per ciascuna città).

Per effettuare questa verifica è stata utilizzata la funzione "table" per visualizzare le frequenze assolute e successivamente, come ulteriore verifica, il calcolo dell'indice Gini.

Task 7 – Calcolo delle probabilità

Il calcolo di probabilità per le variabili city, month e month rispetto ad year è stato quello di andare a plottare le frequenze relative per ciascuna variabile. Essendo variabili discrete si è proceduto con dei grafici a barre al fine di visualizzarle le frequenze relative. L'outcomes del plot di entrambe le variabili è stato che: la variabile city presenta una frequenza relativa distribuita equamente per ciascuna modalità, essendoci nel dataset 4 città, la percentuale con cui si possa ritrovare la città "Beaumont" da un'estrazione casuale è pari al 25%, mentre per la variabile month, la distribuzione si comporta allo stesso modo, ma avendo 12 modalità, la probabilità di estrarre il mese di luglio è pari al 8% (vedasi grafici di seguito).

Infine, per il calcolo della probabilità con cui si possa estrarre il mese di dicembre del 2012 è stato utilizzato lo stesso approccio. Si è proceduto con il calcolo delle frequenze relative delle variabili month e year insieme, per poi passarle a plottare sia sotto forma di boxplot che barplot al fine di valutare la distribuzione delle due variabili.

La variabile month ha avuto come outcome di essere equamente distribuita secondo le sue modalità lungo gli anni del dataset e che la probabilità con cui possa uscire il mese di dicembre del 2012 da un'estrazione casuale è pari a circa 1,7%.

Task 8 – Colonna prezzo medio

Il ragionamento per l'aggiunta della nuova colonna che tracci il prezzo medio per ciascun mese del dataset è stato quello di considerare la colonna "sales" e "volume" che corrispondono al numero totale di vendite e il valore totale delle vendite in milioni di dollari.

Mettendo a rapporto il volume, cioè il valore totale delle vendita espressa in milioni di dollari, su il numero totale di vendite, è possibile ottenere il prezzo medio per ciascun mese. In particolare, essendo volume espresso in milioni di dollari, si è ritenuto opportuno moltiplicarlo per un milione al fine di calcolare il prezzo medio in dollari.

Task 9 – Efficacia delle vendite

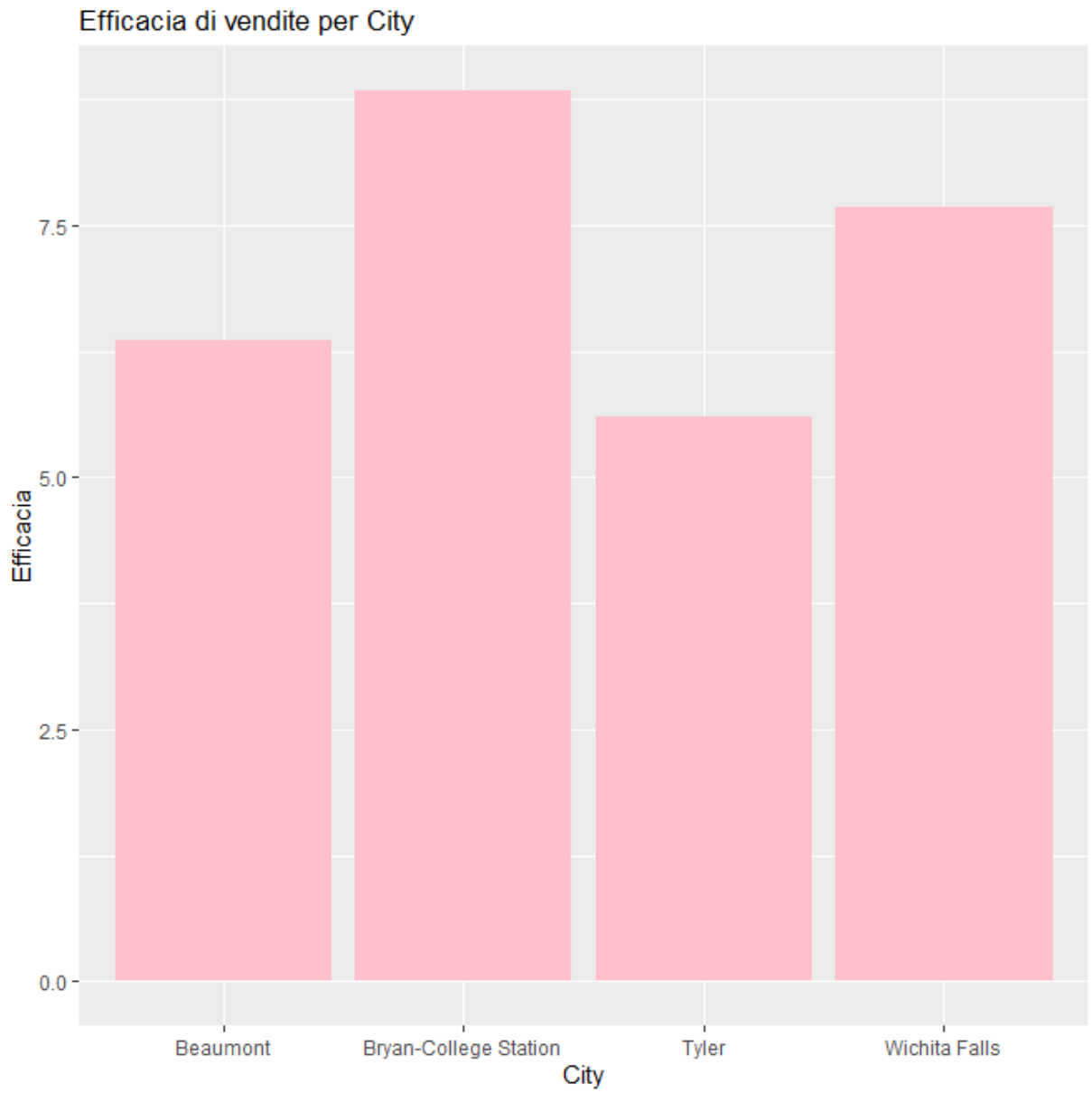
L'efficacia delle vendite è stata calcolata sulla base del fatto che si conoscono le vendite di immobili effettuate e gli immobili ancora non venduti. Il rapporto tra le vendite effettuate e gli immobili ancora non venduti danno l'informazione di quanto sono state efficaci le vendite in ciascuna città per ogni mese degli anni dal 2012 al 2014.

Una volta calcolata l'efficacia, come descritto sopra, si è proceduto anche alla valutazione di dove si è riscontrata una maggiore efficacia rispetto alle città e gli anni, andando a plottare l'efficacia in relazione alle variabili city e year. Si ritenuto opportuno creare un aggregazione dell'efficacia per città, mese ed anno per passare poi alla costruzione di un grafico geom_bar suddiviso per ogni città, andando così a valutare l'efficacia delle vendite in maniera distinta per ogni città, mese e anno.

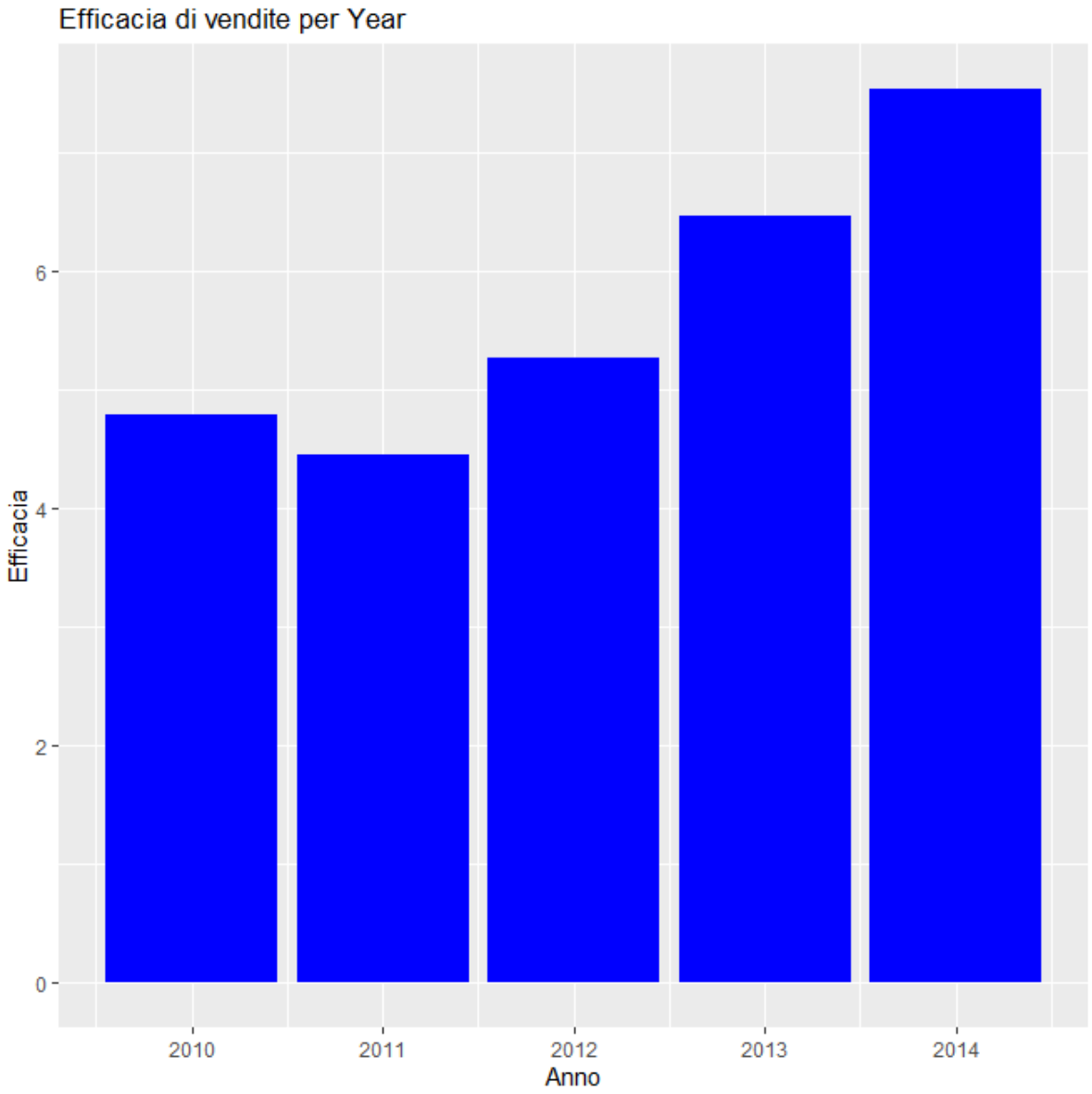
Successivamente, si è pensato di realizzare un grafico 3 dimensionale al fine di riportare in quale mese dell'anno e di quale città si è riscontrata una percentuale di vendite maggiori e quindi un'efficacia di vendita superiore a tutti i dati.

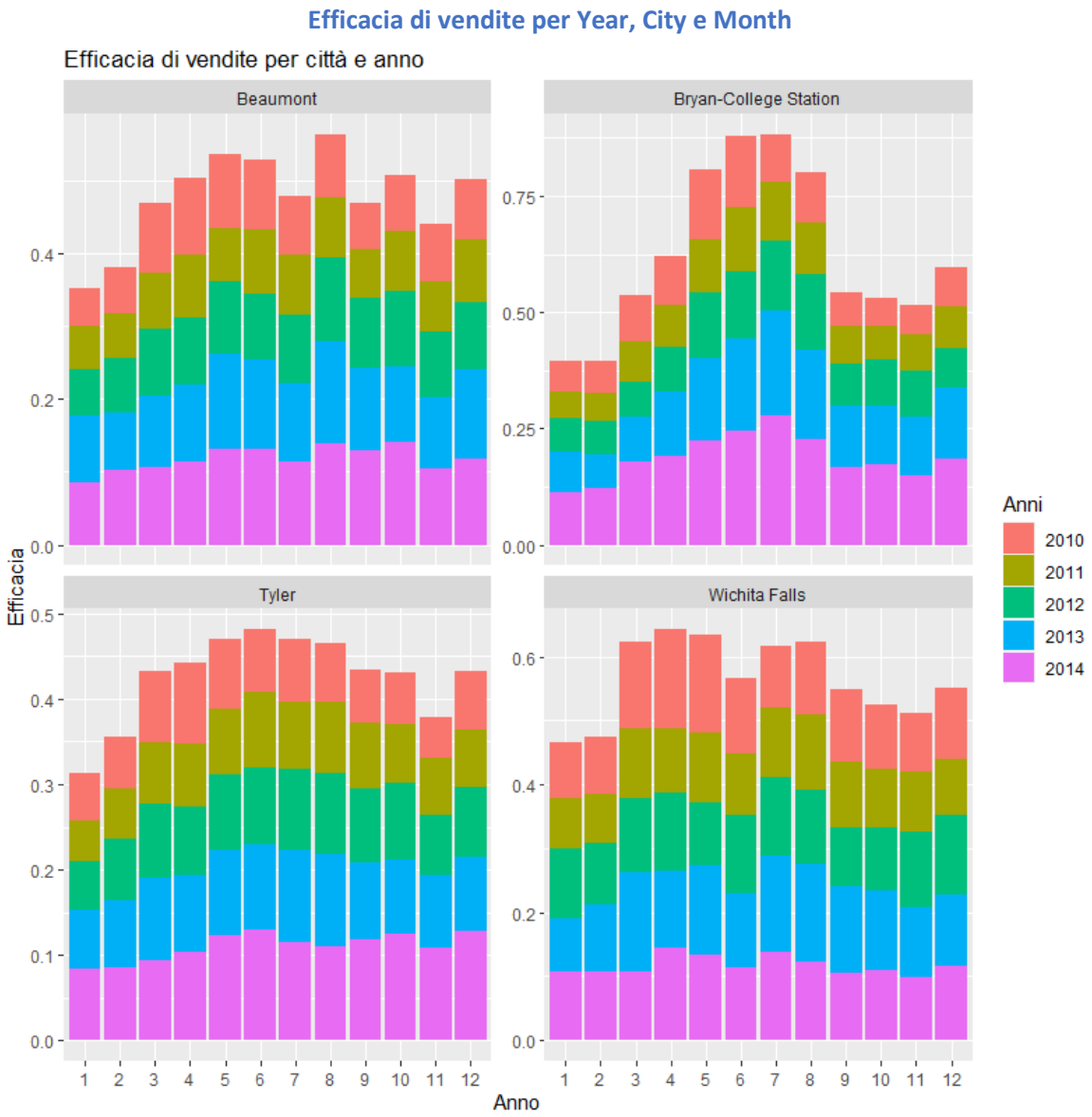
Gli outcomes sono stati che l'efficacia maggiore di vendite è stata riscontrata nella città di "Bryan-College Station" e nell'anno 2014. In particolare, le vendite hanno avuto un maggior successo nel mese di Luglio dell'anno 2014 per la città "Bryan-College Station".

Efficacia di vendite per City



Efficacia di vendite per Year





Task 10 – Summary delle variabili sales, volume e listings

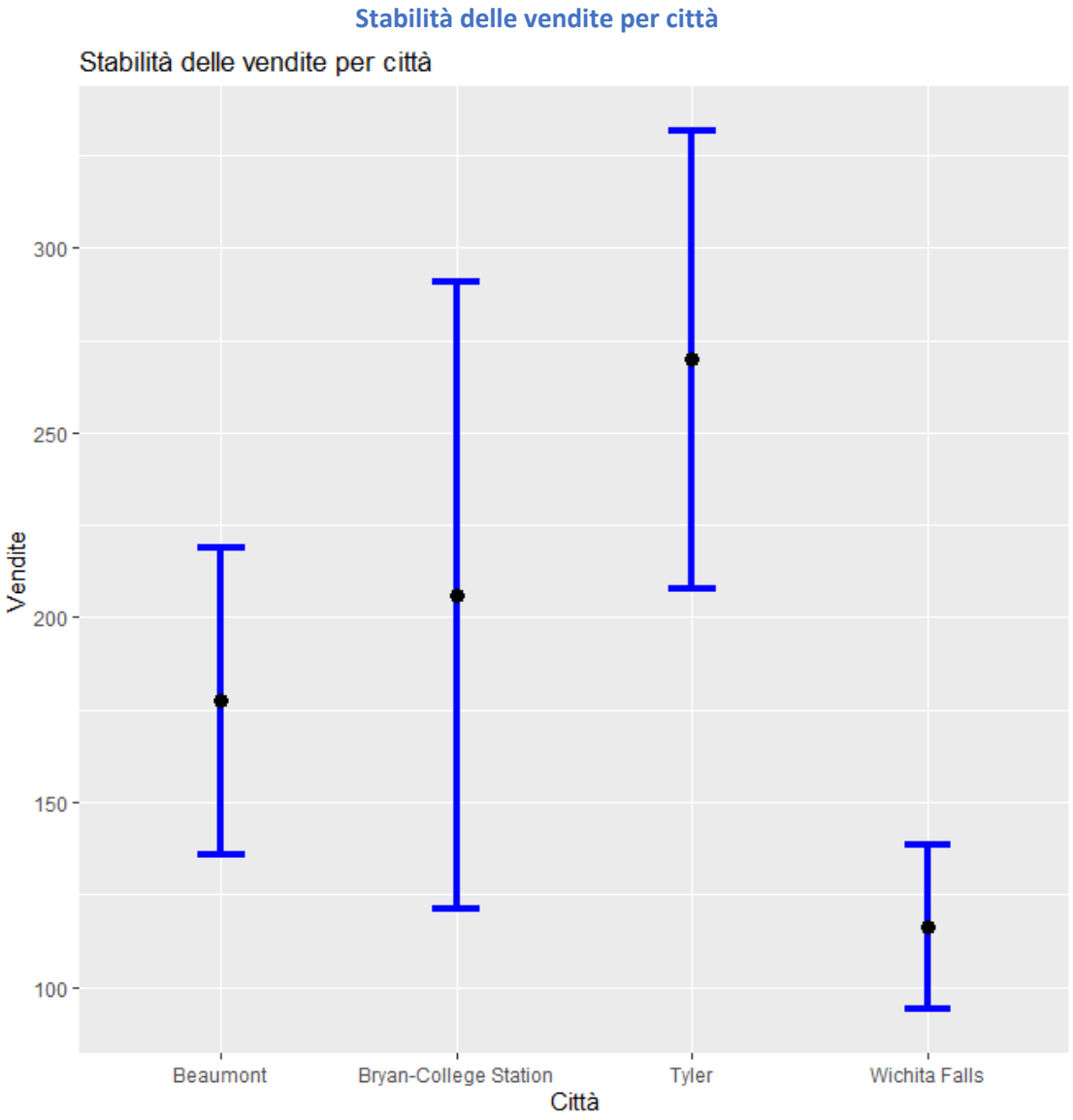
Le variabili scelte per la creazione dei summarise tramite la libreria "dplyr" sono state: sales, volume e listings. L'obiettivo è stato quello di valutarne media e deviazione standard per trarre delle conclusioni su quello che è stato l'andamento delle vendite, dollari e annunci ancora attivi.

Per questo motivo, si è passati alla creazione dei grafici per alcuni dei summarise creati per trarre le conclusioni descritte nell'obiettivo.

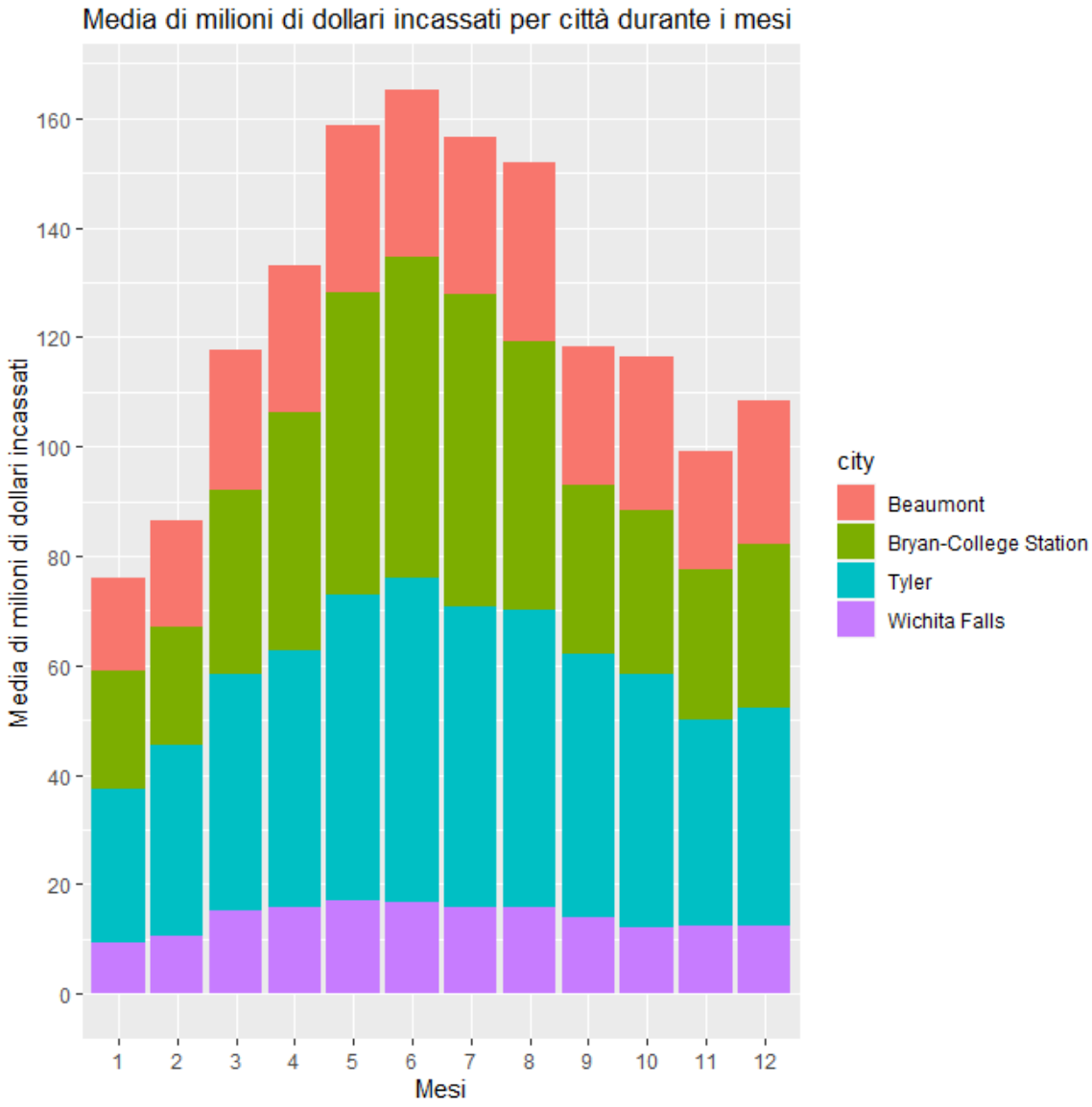
Il primo grafico a barre (geom_errorbar) in cui si è andato a mettere in relazione la media e la deviazione standard delle vendite per ogni città è stato fatto al fine di comprendere in quale città mediamente sono stati venduti maggiormente gli immobili e di quanto esse variano. La città di "Bryan-College Station" è risultata la città con una media di vendite di circa 200 e con una variabilità molto elevata, quindi la città in questione presenta una poca stabilità nelle vendite rispetto alla media, a differenza della città di "Wichita Falls" che presenta una stabilià maggiore con una media di vendite nettamente inferiore.

Successivamente, si è optato per creare un grafico a barre per valutare in quali mesi dell'anno mediamente gli immobili producono un profitto maggiore e in quale città. Per questo si è messo in relazione la media di milioni di dollari con i mesi degli anni e le corrispondenti città. Il risultato è stato che nel mese di "Giugno" la media di dollari incassati dalle vendite è stato maggiore rispetto a tutti gli altri, e in particolare, gli incassi maggiori sono stati riscontrati nelle città di "Tyler" e "Bryan-College Station".

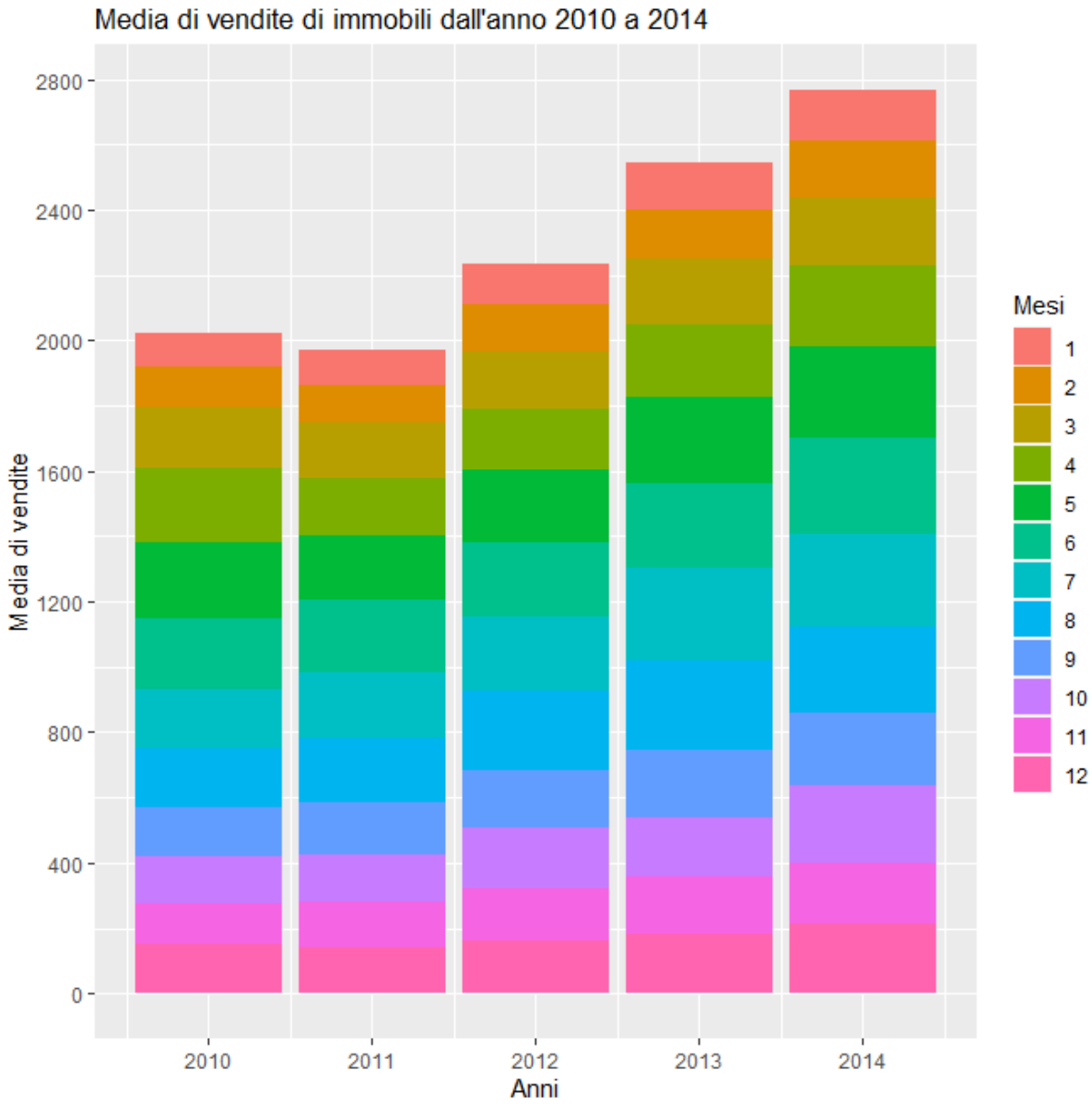
Infine, si è optato di creare un grafico a barre sovrapposte per valutare in quali mesi degli anni dal 2010 e 2014, riportarti dal dataset, si sono verificate maggiormente in medie le vendite. Per questo si è messo in relazione la media delle vendite con gli anni e mesi. Il risultato è stato che dall'anno 2011 all'anno 2014, la vendita di immobili è aumentata e in particolare, nell'anno 2014 sono stati venduti mediamente la maggior quantità di immobili.



Media di milioni di dollari incassati per città durante gli anni



Media di vendite di immobili dall’anno 2010 al 2014



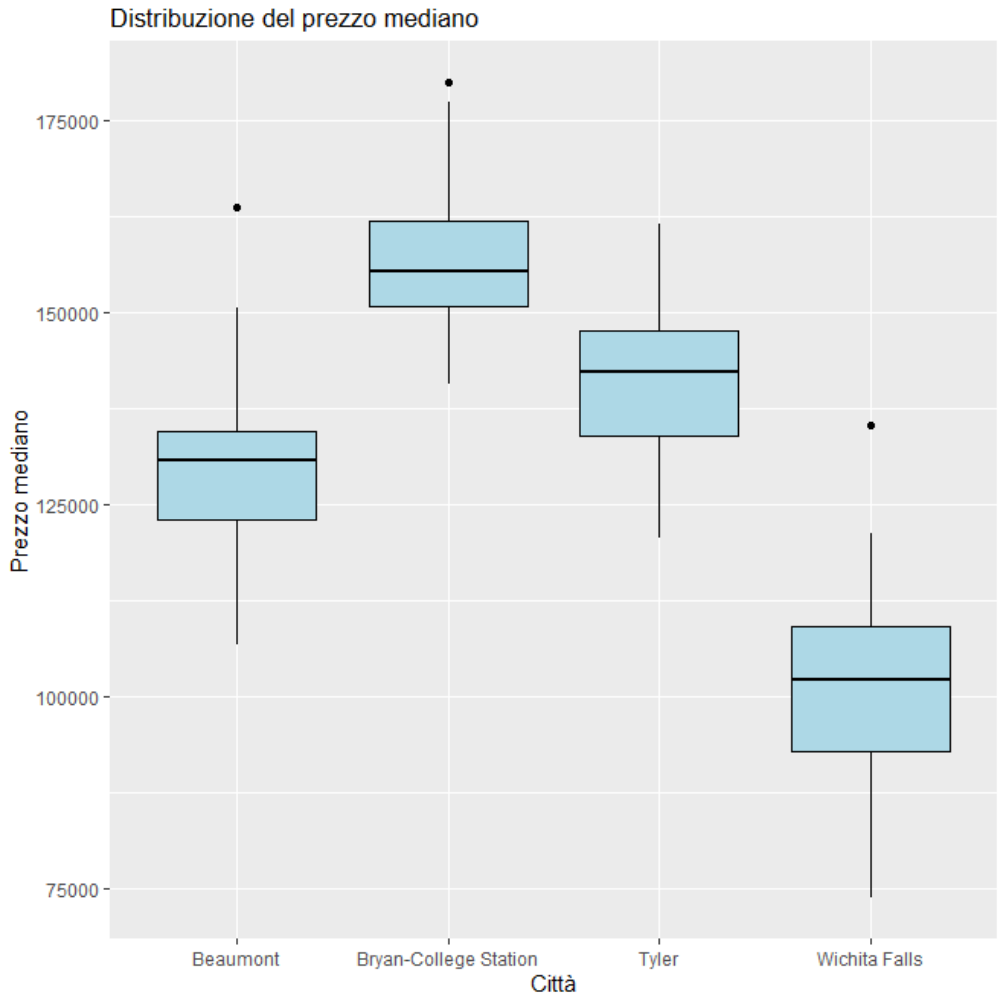
Task 11 – Boxplot median_price tra le città

Il confronto del prezzo mediano tra le città e tra le città in relazione con gli anni è stato eseguito attraverso i grafici boxplot che riassumono alcuni indici di posizione e variabilità con cui poter trarre delle conclusioni sulla distribuzione della variabile median_price.

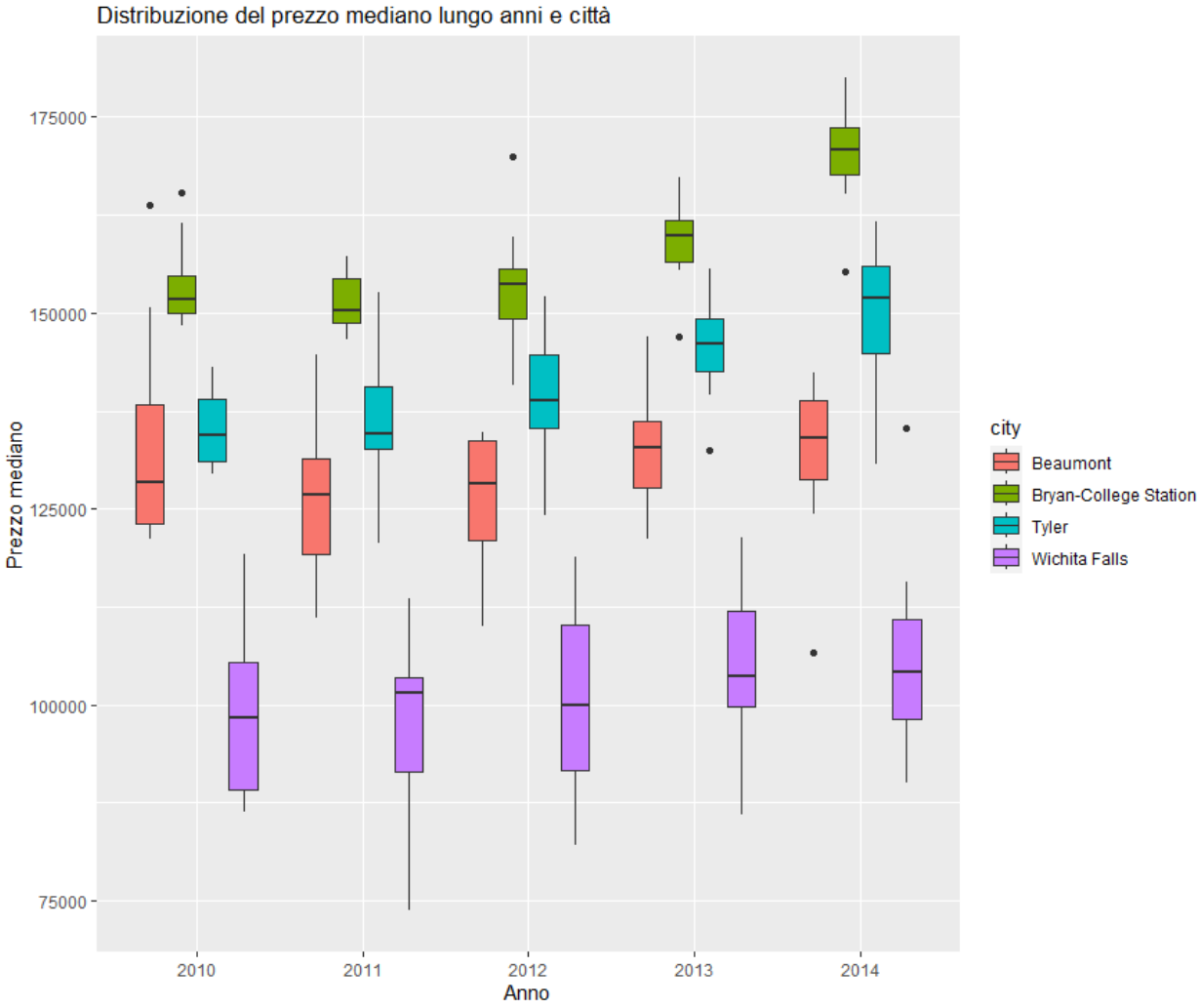
Dal confronto dei boxplot del prezzo mediano per singola città si può notare per la città "Wichita Falls" come la variabilità del prezzo mediano è elevata rispetto alle altre città nonostante il prezzo mediano presenta un range molto inferiore rispetto alle altre città. Nel caso opposto invece, è possibile notare come la città "Bryan-College Station" presenta una variabilità inferiore rispetto alla città "Wichita Falls" e il suo range interquartile risulta più piccolo denotando una maggiore stabilità del prezzo mediano lungo la città in questione.

Dal confronto dei boxplot del prezzo mediano per singola città condizionatamente agli anni, considerati come fattori, è possibile dedurre che durante gli anni dal 2010 e 2014, per la città "Wichita Falls" il prezzo mediano è variato di molto presentando anche un valore di outline nell'anno 2014. Questo può significare che nella città di "Wichita Falls" il prezzo mediano non è molto stabile e varia di molto rispetto alle altre città. Nel caso opposto, invece, è possibile notare nella città di "Bryan-College Station" come il prezzo mediano, dal 2010 al 2014, risulta più o meno costante e non presenta una variabilità elevata. Inoltre, è possibile notare come nel 2014, il prezzo mediano risulta essere aumentato di parecchio rispetto agli anni precedenti.

Distribuzione del prezzo mediano lungo le città



Confronto della variabilità del prezzo mediano lungo anni e città

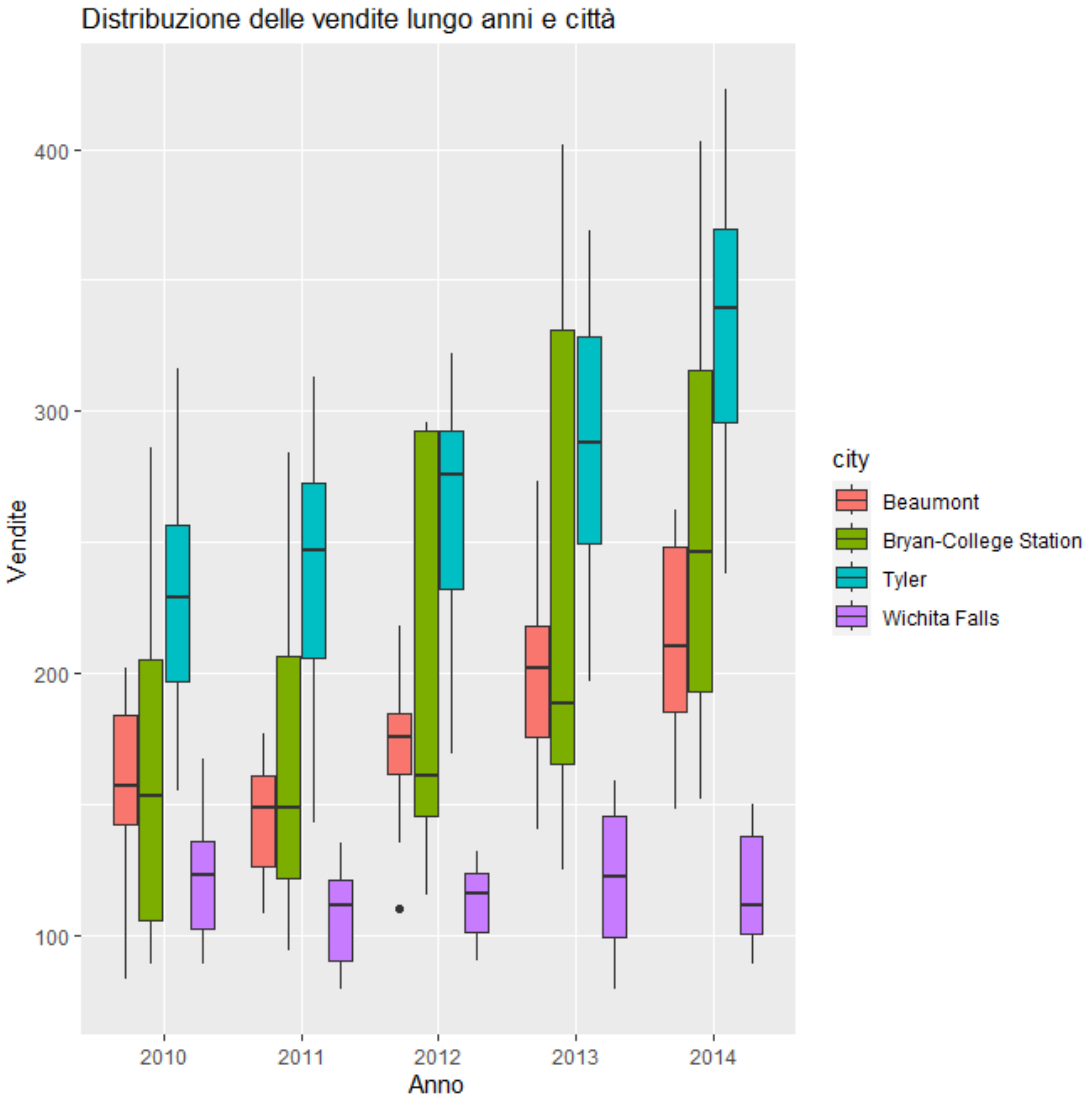


Task 12 – Boxplot sales tra le città

La valutazione delle vendite effettuate durante gli anni dal 2010 al 2014 è stata effettuata attraverso la comparazione dei boxplot delle vendite condizionatamente ad ogni singola città per ciascun anno dal 2010 al 2014. Successivamente, si è proceduto alla creazione di un grafico a barre sovrapposte con la stessa relazione tra le variabili sales, city e year.

Dai boxplot, si è potuto notare come la variabilità delle vendite nella city “Bryan-College Station” durante gli anni è aumentata nel tempo e, in particolare, nell’anno 2013, sono oscillate di molto rispetto agli anni precedenti e successivi. La città che ha riscontrato maggior successo nelle vendite è “Tyler”, la quale negli anni presenta una continua crescita del proprio range arrivando anche ad una quota di vendite superiore a 400. Infine, la città che nel tempo è restata piò o meno costante, ma che nel 2014 ha presentato un calo di vendite è stata “Wichita Falls”.

Confronto della variabilità del prezzo mediano lungo anni e città



Task 13 – Grafico a barre sovrapposte per vendite su anni e mesi

Il grafico a barre sovrapposte per la valutazione delle vendite di immobili condizionatamente agli anni, mesi e città è stato realizzato dividendo per ciascuna città un grafico nel quale si è andato a plottare per ogni anno, il numero di vendite per ogni mese con una scala di colorazione del blu per evidenziare e marcare i diversi mesi.

Dal grafico si evince per ciascuna città:

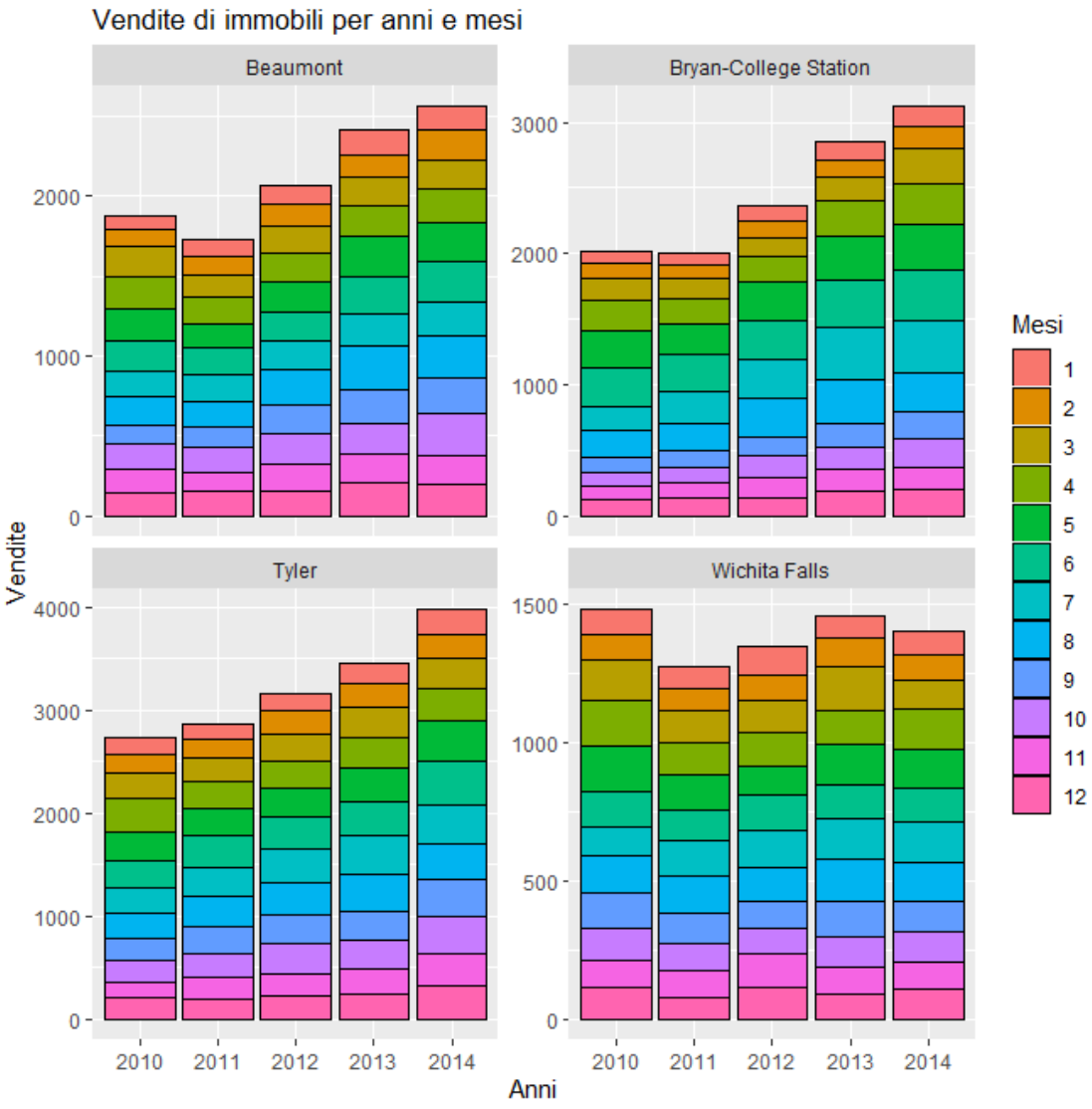
- **Beaumont**: le vendite sono calare nell'anno 2011 rispetto a tutti gli altri e nel 2014 si è registrato un aumento delle vendite
- **Bryan-College Station**: le vendite sono state più o meno costanti e in continua crescita durante gli anni, toccato il picco nel 2014. In particolare, le vendite maggiori si sono registrate durante il periodo di Giugno e Luglio.

- **Tyler**: le vendite sono state in continua crescita e infatti è la città in cui si sono venduti maggiormente gli immobili. In particolare, si può notare come le vendite sono state costanti anche durante tutti i mesi rispettivamente per ciascun anno
- **Wichita Falls**: le vendite hanno oscillato abbastanza durante gli anni toccando il picco di vendite nel 2010 per poi abbassarsi nel 2011. In particolare, si può notare come le vendite sono state maggiori durante i mesi di Maggio e Luglio.

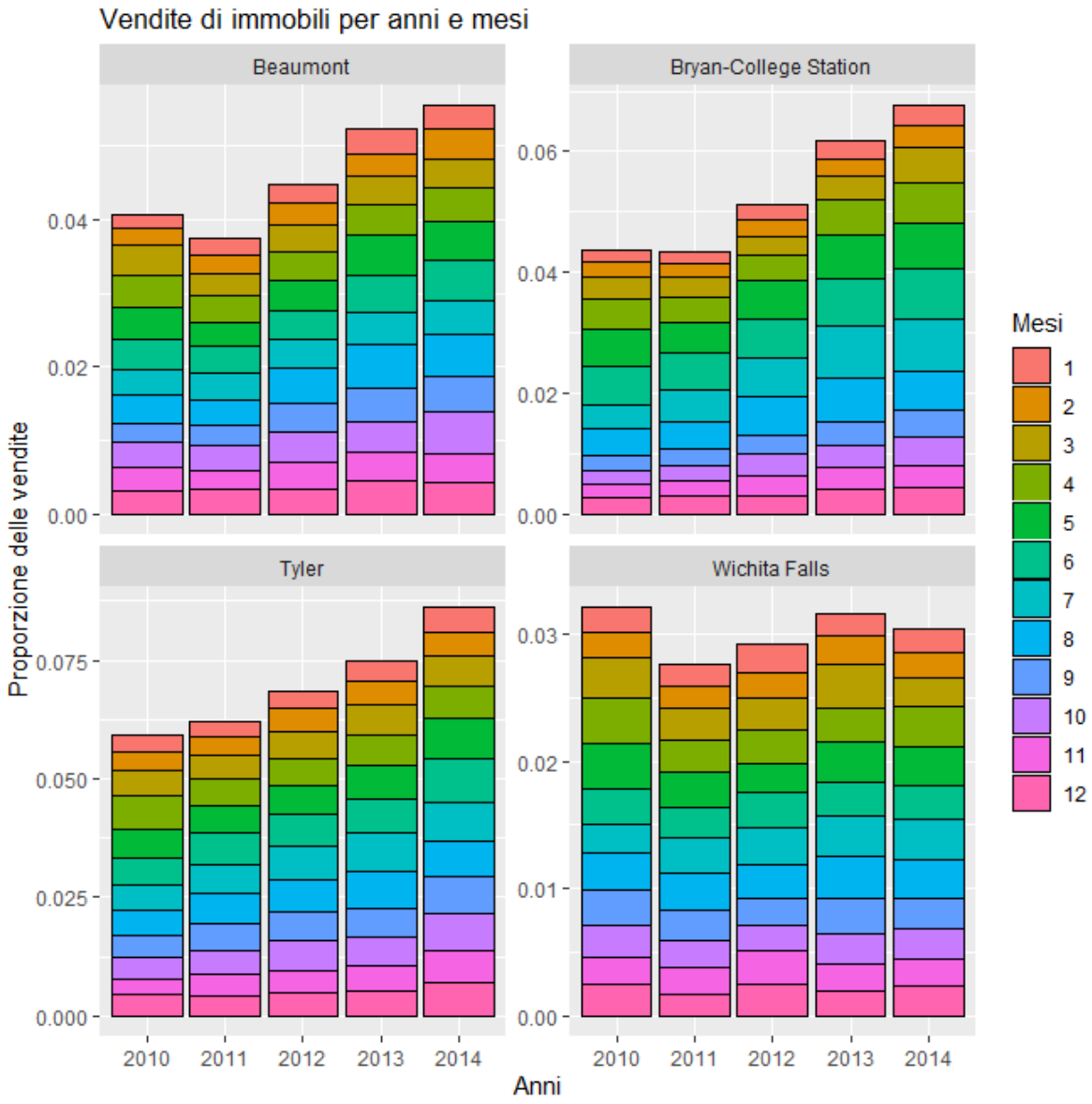
Dal grafico a barre normalizzato è possibile trarre conclusioni sulle vendite condizionatamente a ciascun mese ed anno, rispetto al totale delle vendite su tutte le città, in tutti gli anni e mesi.

A conferma delle evidenze tratte dal grafico a barre non normalizzato, è possibile notare come per le città di "Bryan-College Station" e "Tyler", le vendite rappresentano rispettivamente circa il 6% e 8% delle vendite totali nell'anno 2014, quindi rappresentano le città che hanno riscontrato un maggior numero di vendite in quell'anno; mentre per le città di "Beaumont" e "Wichita Falls", le vendite durante gli anni dal 2010 al 2014, hanno rappresentato rispettivamente circa il 3% e 5%.

Vendite degli immobili lungo gli anni e mesi nelle città



Vendite degli immobili lungo gli anni e mesi nelle città con grafico a barre normalizzato



Task 14 – Line chart

La realizzazione del line chart è stata eseguita sulla variabile “volume” del dataset che rappresenta in milioni di dollari, i soldi incassati dalle vendite degli immobili. Si è pensato di andare a plottare lungo ciascun anno, il volume di dollari incassati per ciascun mese ed evidenziare attraverso i colori le varie città.

Dal grafico si evince per ciascuna città

- **Beaumont**: i milioni di dollari sono oscillati durante gli anni dal 2010 al 2014 intorno ai 20 – 40 milioni di dollari, toccando il suo massimo nel 2013 con circa 41 milioni di dollari incassati per le vendite di immobili, nel mese di Agosto.

- **Bryan-College Station**: i milioni di dollari incassati hanno raggiunto la quota di 80 milioni nell'anno 2014 nel mese di Luglio. Per lo stesso mese e anche quello di Agosto, negli anni 2012 e 2013, si sono raggiunti picchi di incassi rispettivamente nell'intorno di 50 – 65 milioni.
- **Tyler**: i milioni di dollari incassati maggiormente sono stati durante gli anni 2014 e 2013, in cui si sono raggiunti picchi di 60 – 80 milioni di dollari, principalmente nei mesi di Giugno e Luglio per i rispettivi anni.
- **Wichita Falls**: i milioni di dollari incassati sono stati relativamente bassi rispetto a tutte le altre città, solamente nell'anno 2010 si sono riscontrati incassi maggiori durante i mesi da Marzo a Giugno.

In definitiva, è possibile concludere che la maggior parte degli incassi sono stati effettuati durante i periodi primaverili/estivi per ciascuna città, in particolare, l'anno 2014 ha rappresentato l'anno con il maggior successo di incassi.

Il tutto trova corrispondenza con le analisi e conclusioni tratte dai grafici che riportano le vendite per ciascuna città e periodo storico.

Milioni di dollari incassati per anni, mesi e città