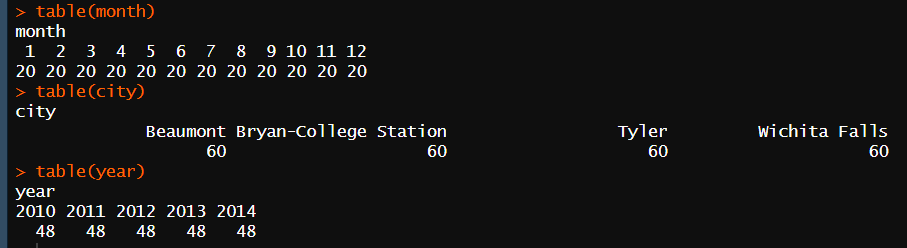
**Progetto statistica descrittiva**

Vittorio D’Onofrio

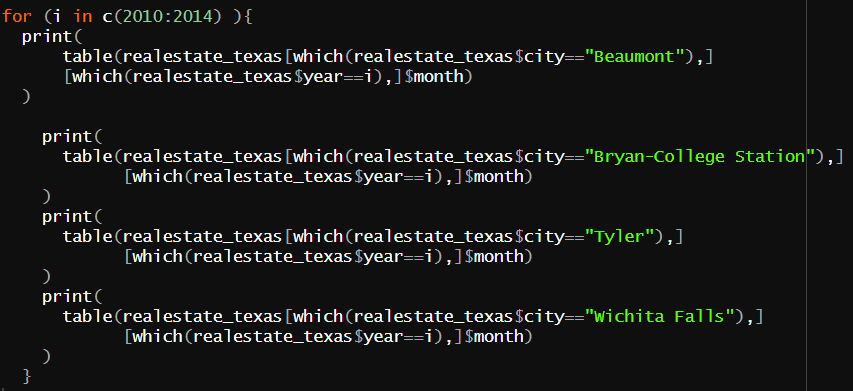
Quesito 1) Il dataset è stato scaricato e poi importato in R con la funzione “read.csv”.

Quesito 2)

* La variabile sales è quantitativa discreta perché rappresenta il conteggio delle vendite.
* La variabile listings è quantitativa discreta perché rappresenta il conteggio degli annunci.
* La variabile months\_inventory è quantitativa continua perché assume valori su scala continua e può essere suddivisa in intervalli più piccoli.
* La variabile year è quantitativa su scala di intervalli.
* La variabile city è qualitativa nominale perché le modalità si possono confrontare solo in termini di uguaglianza.
* La variabile month è qualitativa anche se espressa in numeri.
* La variabile volume è quantitativa continua, può essere suddivisa in intervalli più piccoli; ad esempio si può pensare di esprimere il prezzo in dollari, milioni di dollari o migliaia di dollari.
* La variabile median\_price è quantitativa continua, rappresenta un prezzo che può essere scritto in dollari, migliaia di dollari o milioni di dollari ad esempio.

Quesito 3) Si osserva che per le variabili “year”, “month” e “city”, tutte le modalità presentano la stessa frequenza assoluta:

Si intuisce che ogni città ha esattamente una osservazione al mese da Gennaio 2010 a Dicembre 2014. Sarà vero? Cerco di provare l’intuizione usando R.



Con queste righe di codice si trova quante osservazioni ha una città fissata al mese per un anno fissato.

Il risultato è uguale per ogni iterazione del ciclo ed è uguale al seguente:

Immagine che contiene testo, schermata, Carattere, tastiera

Descrizione generata automaticamente

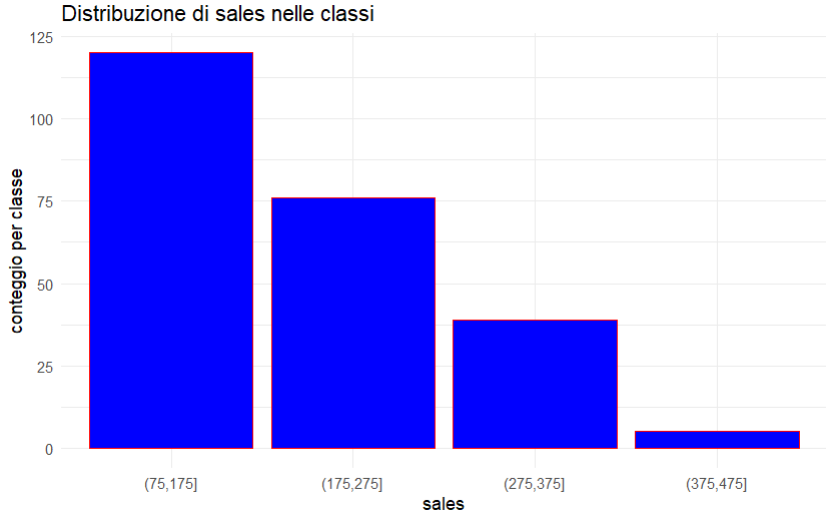
Questo effettivamente dimostra che ogni città ha un’osservazione al mese per ogni mese da Gennaio 2010 a Dicembre 2014. Questa osservazione può essere utile già da sola in quanto prova che non ci sono righe ripetute: ad esempio è escluso il caso in cui una città abbia due righe con lo stesso anno e lo stesso mese e magari le altre variabili diverse(questa cosa se dovesse succedere, per come è la struttura dati, effettivamente sarebbe un errore); l’osservazione esclude questo caso. In più questo fatto semplificherà qualche conto nel quesito 7.

Quesito 4) Per trovare la variabile che ha variabilità più elevata basta trovare la variabile che ha coefficiente di variazione massimo. Per trovare invece quella più asimmetrica basta trovare la variabile con indice di asimmetria massimo. Si osserva, quindi che in entrambi i casi la variabile con maggiore variabilità e più asimmetrica è “volume”.

Quesito 5) La variabile sales è stata divisa in quattro classi, ogni classe ha 100 valori. Di seguito la distribuzione di frequenze e il grafico a barre.

. Immagine che contiene testo, Carattere, schermata, numero

Descrizione generata automaticamente



L’indice di Gini della variabile sales è pari a 0.998379.

Quesito 6) L’indice di eterogeneità di Gini per la variabile city è uguale a 1 perché dalle considerazioni del punto 3 abbiamo che ogni modalità di questa variabile compare lo stesso numero di volte (60 righe per ogni città) e dunque siamo nel caso di eterogeneità massima. Possiamo anche calcolare l’indice tramite la formula, infatti abbiamo che la frequenza assoluta per ognuna delle 4 città è pari a 60, quindi la frequenza relativa di ogni città è pari a 60/240=1/4. Elevando al quadrato, si ha che f\_{j}^{2}=1/16 per ogni j. G=1-1/4=3/4. Dunque G’=( 3/4)/(3/4)=1.

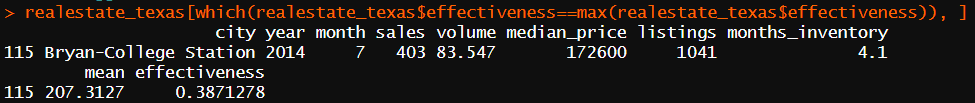
Quesito 7)

* Dalle considerazioni fatte al quesito 3, sappiamo che la città di “Beaumont” compare in 60 righe; le righe totali della tabella sono 240, quindi P(city="Beaumont")= #righe in cui compare “Beaumont”/ #righe totali=60/240=1/4.
* Dalle considerazioni fatte al quesito 3 sulla moda sappiamo che il mese di Luglio compare in 20 righe, le righe totali sono sempre 240, allora P(month=7)= #righe in cui month=7/#righe totali=20/240 = 1/12.
* Partiamo dal fatto che P(month=12 e year=2012)= P(month=12 | year=2012) \*P(year=2012). Ora, sempre dalle considerazioni fatte al quesito 3, fissata una città e un anno c’è un’osservazione al mese. Perciò, visto che le città sono 4 e il numero di righe in cui year=2012 è uguale a 48 (questo sempre dalle considerazioni sulla moda); vale che P(month=12 | year=2012)= 4/48. Inoltre P(year=2012)= 48/240. Dunque P(month=12 e year=2012)=(4/48)\*(48/240)=1/60.
  + In alternativa si può usare il pacchetto dplyr di R e filtrare la tabella per month=12 e year=2012, in questo modo la probabilità cercata è uguale al numero di righe della tabella filtrata diviso il numero di righe della tabella originaria, quindi P(month=12 e year=2012)=4/240=1/60.

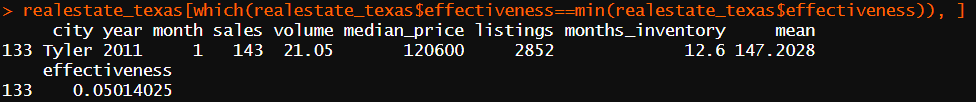
Quesito 8) La variabile richiesta si ottiene come rapporto tra “volume” e “sales”.

Quesito 9) La variabile “effectiveness” dato dal rapporto tra “sales” e “listings” può dare idea di quanto siano stati efficaci gli annunci in un dato mese di una data città nel dato anno. Infatti se effectiveness in una riga è uguale a 1 significa che per quella riga tutti gli annunci si sono trasformati in vendita; più effectiveness è vicino a 1 e più significa che gli annunci per quel mese sono stati efficaci, se invece effectiveness è vicino a 0 significa che gli annunci per quella riga sono stati poco efficaci in quanto pochi annunci si sono trasformati in vendita.

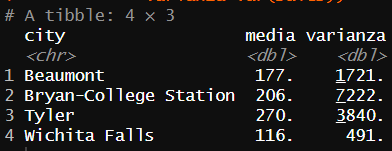
Di questa nuova colonna si può trovare il massimo, il minimo, la media ad esempio.

Tramite la seguente riga di codice si può stabilire che gli annunci più efficaci sono quelli della città di Bryan-College Station nel Luglio del 2014.

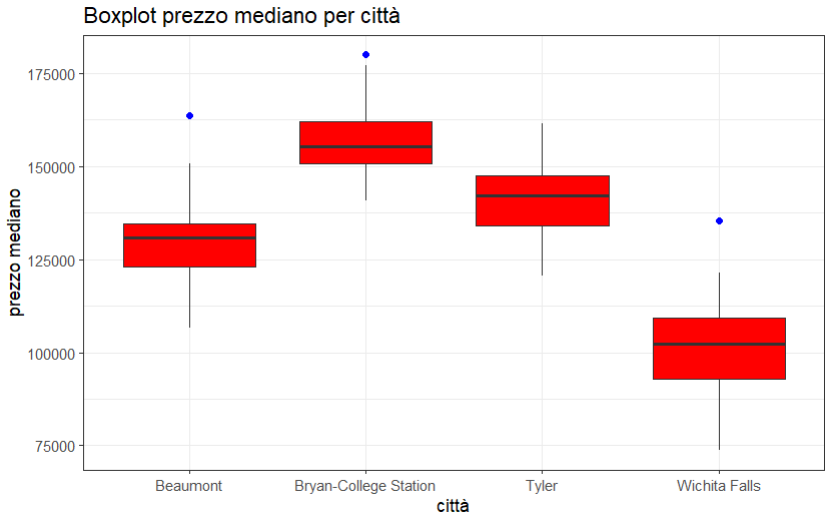
In maniera analoga si può vedere quali sono stati gli annunci meno efficaci:



Quesito 10) Sono state calcolate media e varianza di sales raggruppando prima per città, poi per città e anno. Sono state calcolate media, mediana, massimo, minimo e varianza della variabile sales e poi di median\_price raggruppando per città e mese. Di seguito ad esempio, media e varianza della variabile sales raggruppando per città.



Quesito 11) Di seguito il boxplot utilizzato per confrontare la distribuzione del prezzo mediano tra le varie città.



Dal boxplot possiamo osservare innanzitutto che la distribuzione di median\_price per città ha pochi outliers (gli unici sono i massimi di Beaumont, Bryan-College Station e Whichita Falls); questo fatto lascia pensare che la distribuzione sia “quasi” simmetrica, e che quindi media e mediana siano molto vicine. A questo punto dal boxplot si può fare qualche considerazione sulla varianza. Innanzitutto, l’ampiezza di ogni box rappresenta la distanza interquartile e sicuramente la mediana è compresa in questo range. Quindi, se il box è poco ampio questo vuol dire che il 50% centrale dei valori è vicino alla mediana, viceversa se il box è più ampio ci sarà più dispersione dei valori dalla mediana. Nel nostro caso, abbiamo che media e mediana sono vicine e quindi più è ampio un box e più i valori si discostano dalla media, meno è ampio e meno i valori si discostano dalla media. Dunque, in questo caso più ampio sarà il box e più alto sarà il valore della varianza, meno ampio sarà il box e più bassa sarà la varianza. Infatti, la città con varianza più alta è Whichita Falls e ad occhio dal boxplot sembra la città con il box più ampio. In generale usando il pacchetto dplyr si può creare un summary in cui i valori presenti sembrano confermare le idee venute fuori osservando il boxplot, in particolare il fatto che media e mediana siano vicine.

Quesito 12) Di seguito il boxplot usato per confrontare la distribuzione delle vendite tra città ed anni.

Immagine che contiene testo, diagramma, Diagramma, schermata

Descrizione generata automaticamente

Si osserva che anche in questo caso ci sono pochi outliers nella distribuzione e quindi valgono le considerazioni del quesito 11. Si può vedere dal boxplot che la città di Whichita Falls ha mediamente il numero più basso di vendite ed avendo poca ampiezza i box relativi a questa città, anche la varianza sarà bassa. La città di Tyler ha mediamente un numero crescente di vendite negli anni, questo si osserva notando dove si trova la mediana nei box di questa città. Per la città di Bryan College station si registra il valore più alto della varianza nel 2013. La città di Beaumont invece ha tre outliers nel 2012.

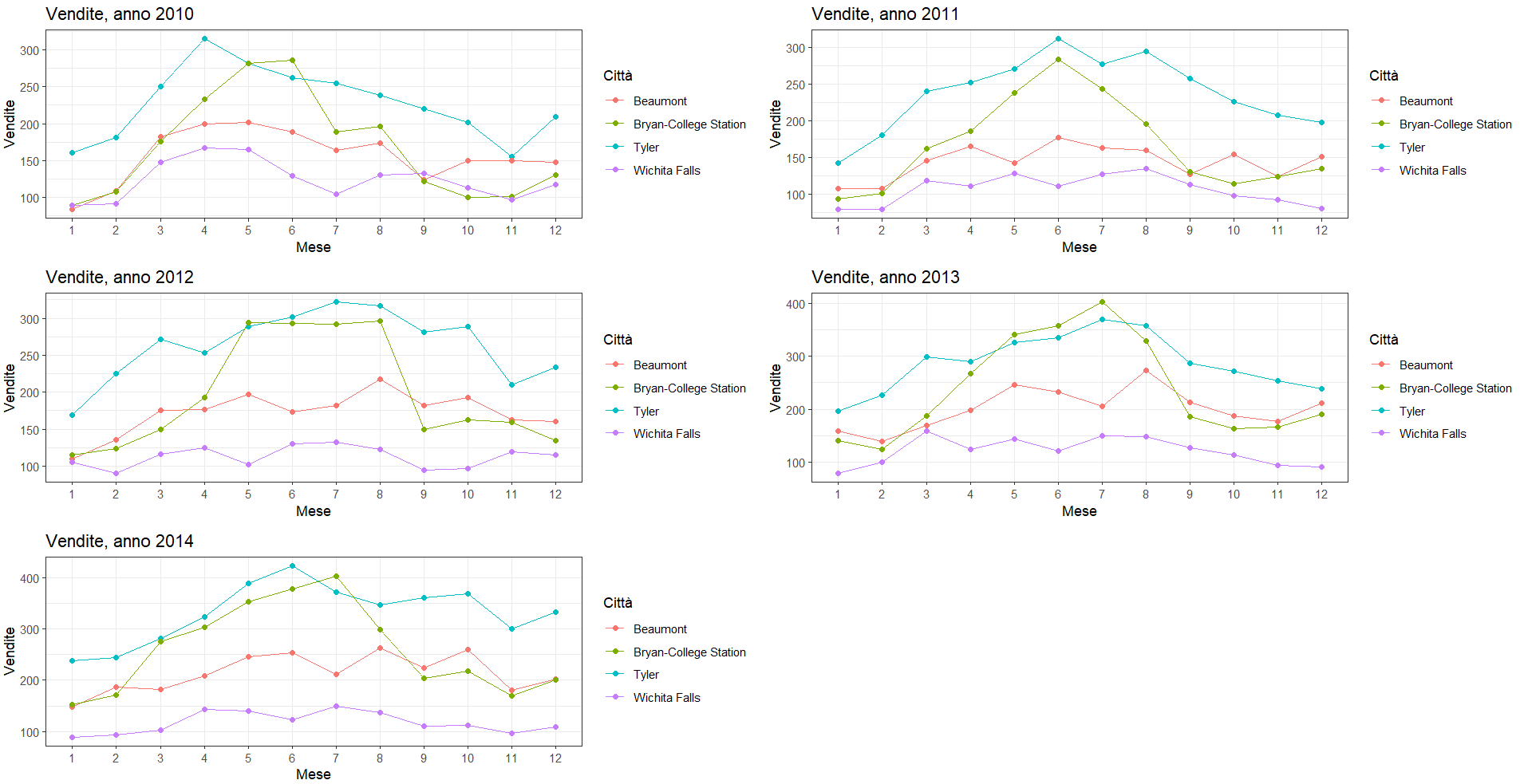
Quesito 13) Di seguito i grafici a barre sovrapposte per confrontare il totale delle vendite nei mesi per ogni anno e per ogni città.

Immagine che contiene schermata, Policromia, testo, diagramma

Descrizione generata automaticamente

Il grafico suggerisce che la città di Tyler presenta vendite maggiori delle altre città nel mese di Dicembre. Con il pacchetto dplyr si può verificare che effettivamente nel mese di Dicembre la città di Tyler è quella che ha registrato più vendite in tutti gli anni.

Quesito 14)Di seguito un line chart della variabile sales per ogni anno.



Si osserva che la città di “Whichita-Falls” ha vendite più basse delle altre città per tutti i mesi di tutti gli anni tranne che per Settembre ed Ottobre del 2010. La città di “Bryan-College station” sembra avere concentrazione di vendite maggiori nei mesi centrali di ogni anno; la distribuzione delle vendite di questa città infatti sembra seguire una Gaussiana centrata nel mese di Giugno.