Progetto fine corso Statistica Descrittiva, Novembre 2021

Tasks:

1-> Importando il dataset in R possiamo constatare che è formato da 240 osservazioni e 8 variabili

2-> Le tipologie di variabili contenute nel dataset sono:

- City: **qualitativa** o categoriale **su scala nominale**
- year: **quantitativa** o numerica **su scala di intervalli**
- month: **quantitativa** o numerica **ordinale**
- sales: **quantitativa** o numerica **discreta**
- volume: **quantitativa** o numerica **continua**
- median_price: **quantitativa** o numerica **continua**
- linstings: **quantitativa** o numerica **discreta**
- months_inventory: **quantitativa** o numerica **continua**

3-> Gli indici di posizione sono:

- media (media aritmetica, media ponderata, media geometrica e media armonica)
- mediana, quartile
- quantile (o percentile)
- moda
- minimo e massimo

Gli indici di variabilità o di dispersione sono:

- campo o intervallo di variazione
- scarto interquartile
- varianza, deviazione standard e coefficiente di variazione

Gli indici di Forma sono:

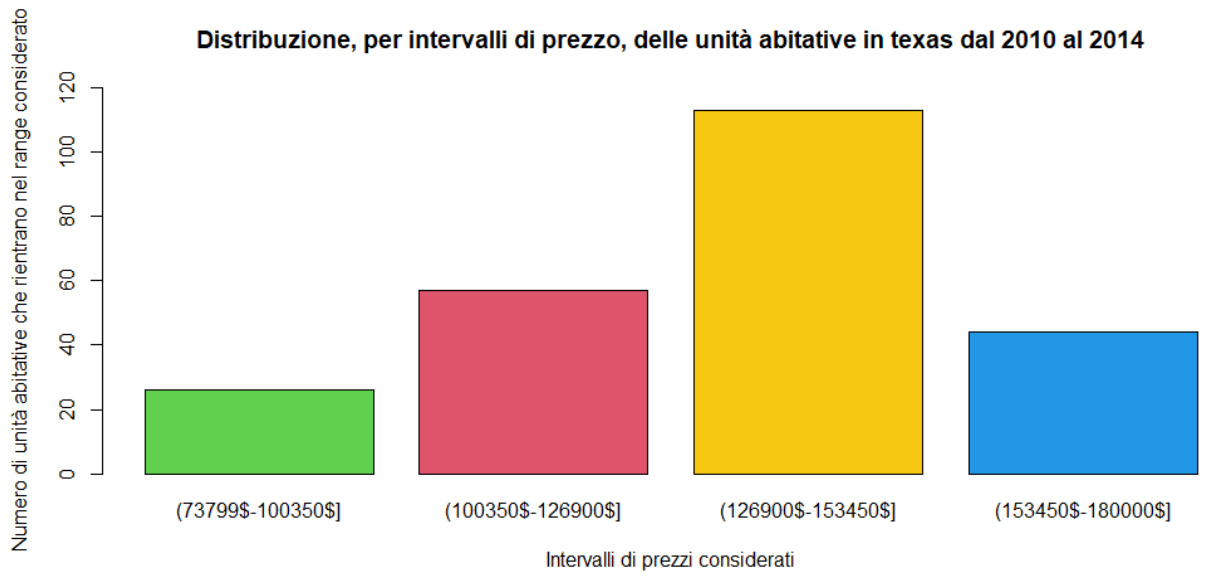
- Asimmetria
- Curtosi

I calcoli degli indici sono presenti sul file File R.

4-> La variabile con variabilità più elevata è “Volume”, come si evince dal calcolo del coefficiente di variazione. Utilizzando l’indice di asimmetria di Fisher riusciamo a determinare che la variabile “Volume” è la più asimmetrica positiva (0.88).

5-> Utilizzando la variabile quantitativa “median_price” ho suddiviso in classi le osservazioni del dataset.

- Sapendo che la variabile median_price ha un valore minimo di 73800$ ed un valore massimo di 180000$, ho creato 4 classi di prezzo crescenti, utilizzando come criterio per l’incremento tra una classe e l’altra l’importo di 26550$.
- Questo valore l’ho ottenuto facendo la differenza tra il valore max e valore min di median_price e poi dividendo quest’ultimo per il numero di classi (4).
- Ho creato la distribuzione di frequenze (distr_freq_median_price_CL) ed il grafico a barre corrispondente. (Plot 1)
- L’indice di Gini per la distr_freq_median_price_CL è (G’=0.753).



Plot 1

6-> Indice di Gini per la variabile City è 1 (G'=1)

7-> La probabilità che presa una riga a caso questa riporti:

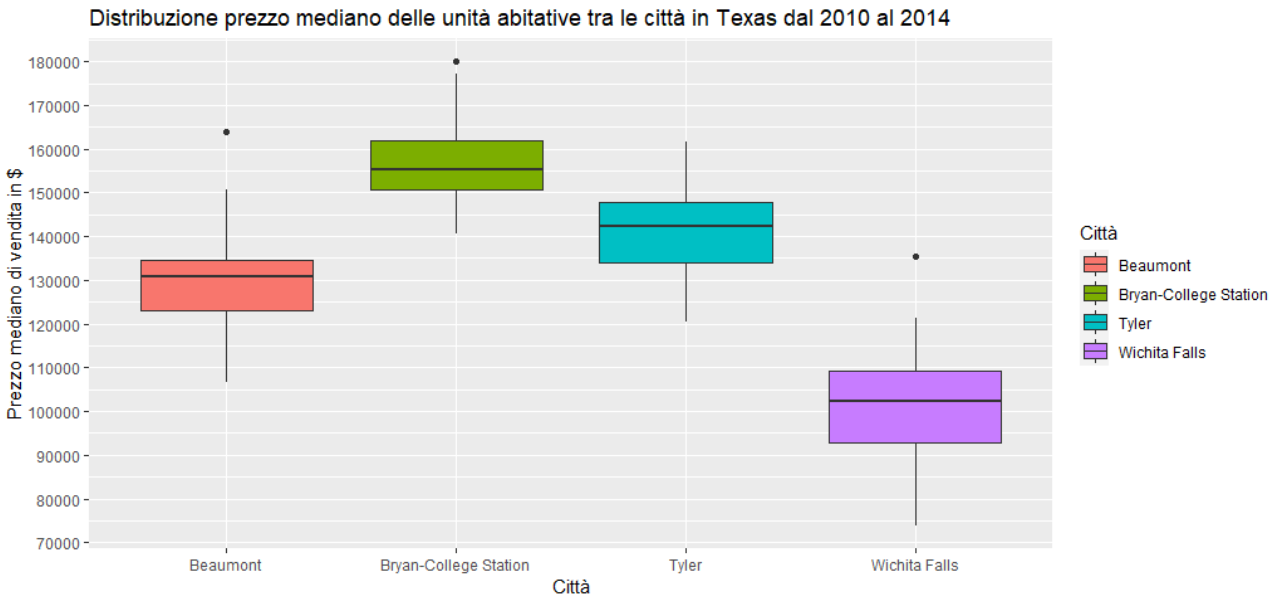
- alla città di Beaumont è del 25%
- il mese di Luglio è del 8.33%
- il mese di Dicembre e l'anno 2012 è del 1.67%

8-> Ho creato una nuova colonna che indica il prezzo medio ottenuta con una piccola funzione scritta ad hoc. Per ottenere la media del prezzo ho utilizzato i valori contenuti nelle variabili "volume", dopo aver convertito il dato da milioni di $ a $, e i dati contenuti nella variabile "sales".

9-> La colonna creata, dal nome "vendite_giornaliere", contiene il valore che indica la stima di vendite giornaliere al ritmo delle vendite attuali. Successivamente ho sommato i valori contenuti nella colonna "vendite_giornaliere", raggruppando i dati per città. Pertanto la città di Tyler ha gli annunci di vendita più efficaci.

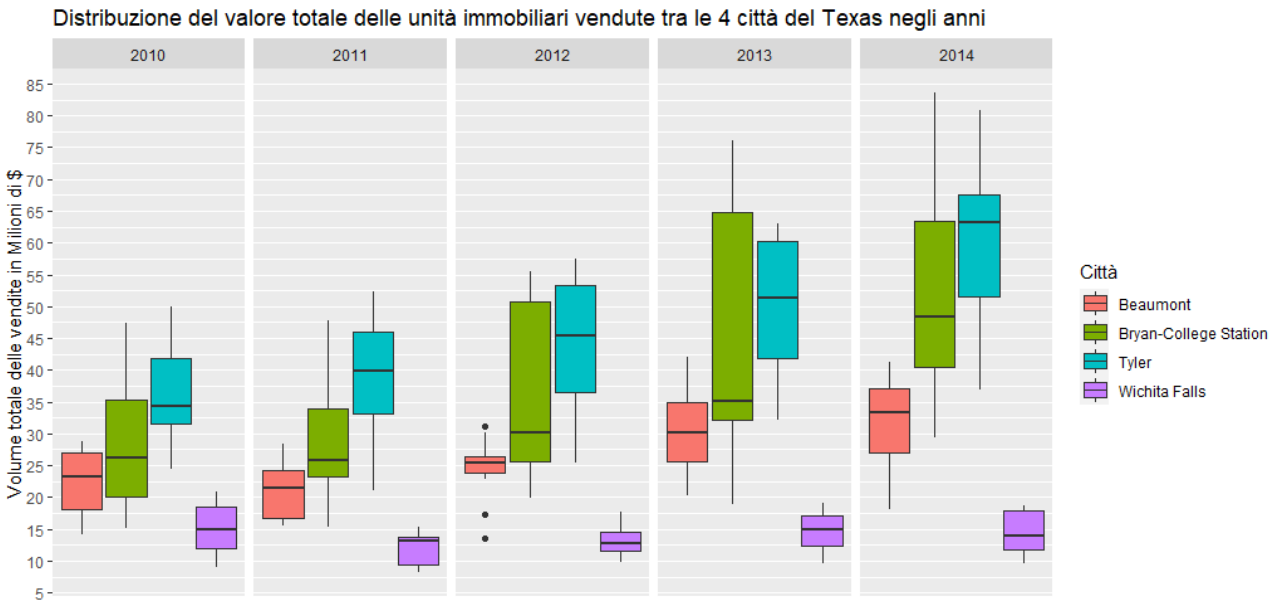
10-> Ho estrapolato media e deviazione standard della variabile prezzo mediano, condizionatamente alle variabili città, mesi ed anni.

10.1-> Come possiamo notare dalla distribuzione del prezzo mediano tra le varie città all'interno del boxplot (Plot 2), troviamo il prezzo mediano più alto nella città di Bryan-Collage station, a seguire Tyler, Beamount e Wichita Falls.



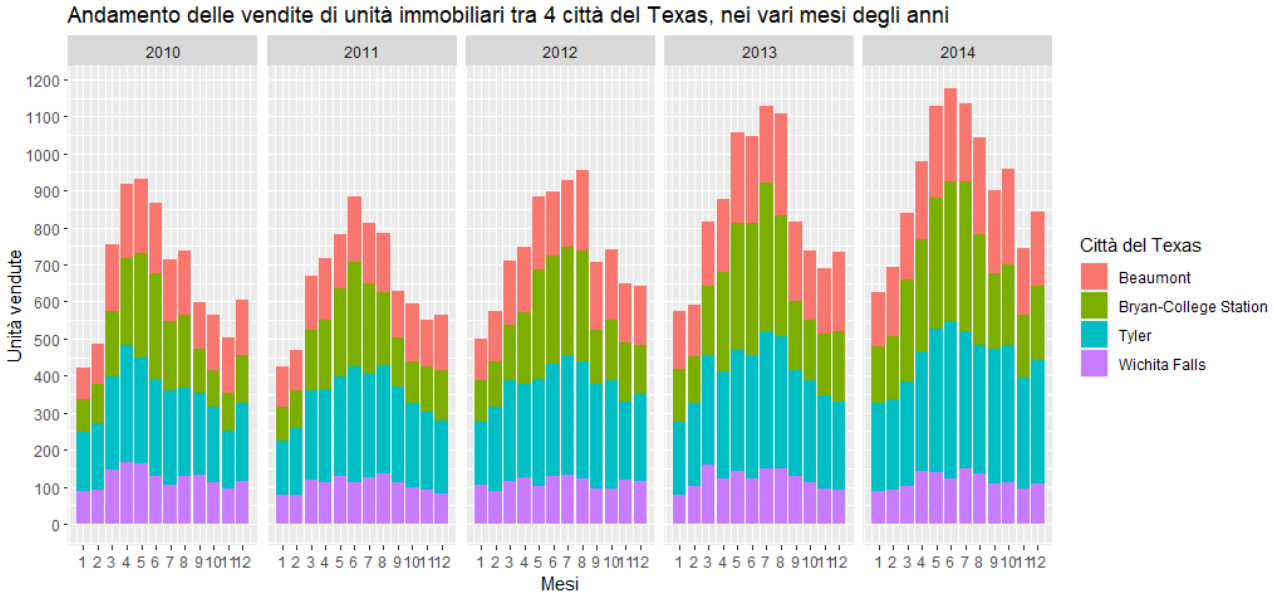
Plot 2

10.2-> Dal grafico creato (Plot 3) è evidente che la distribuzione del valore totale delle vendite tra le varie città è a favore di Tyler, seguita da Bryan-Collage Station, Beamount e Witcha Falls. Questa condizione è costante in tutti gli intervalli considerati(anni). Nel corso degli anni, il volume totale delle vendite, aumenta per tutte le città.

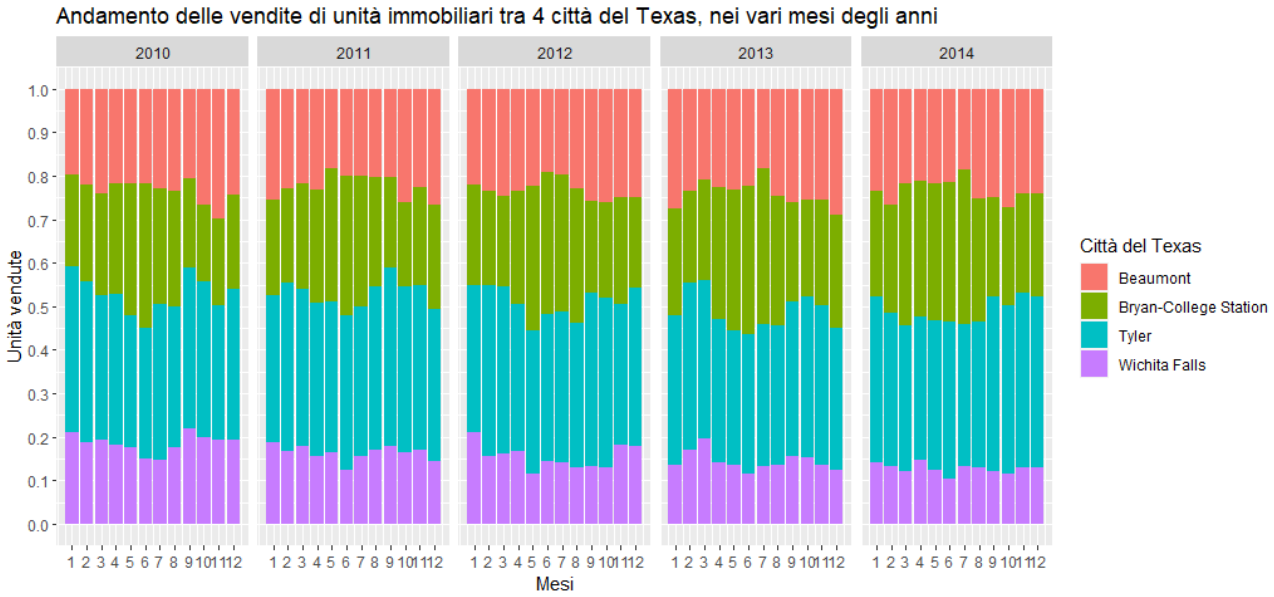


Plot 3

10.3-> Ho creato un grafico a barre sovrapposte (Plot 4) ed uno a barre normalizzato (Plot 5) per confrontare il totale delle vendite di unità abitative delle 4 città nei vari mesi degli anni presi in esame. Dai grafici ho notato che tendenzialmente, le vendite di unità abitative, si concentrano tra il secondo e terzo trimestre in ogni anno considerato.



Plot 4



Plot 5

10.4-> Per la creazione di un line chart che confrontasse le informazioni della variabile prezzo mediano con città e periodi storici ho incontrato qualche difficoltà, tant'è che ho avanzato per gradi.

1. Creazione line chart per singola città e singolo anno
2. Creazione line chart per tutte le città e singolo anno
3. Infine, line chart per tutte le città e l'intero periodo storico

Nel file R maggiori dettagli.