ANALISI ESPLORATIVA DEL MERCATO IMMOBILIARE DEL TEXAS

Precondizioni: Imposto cartella di lavoro

```
getwd()
setwd("/Users/veronicamandelli/Desktop/project_Texas")
```

```
help(nomefunzione)
?nomefunzione
```

```
# carico tutte le librerie
#install.packages("ggplot")
library(ggplot2)
#install.packages("magrittr")
library(magrittr)
#install.packages("dplyr")
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
#install.packages("moments")
library(moments)
```

1. Importa il dataset "Real Estate Texas.csv", contenente dei dati riguardanti le vendite di immobili in Texas.

```
dati <- read.csv("Real Estate Texas.csv", sep = ",", encoding = 'latin1')
#estraggo le colonne da un data frame (per non usare simbolo $)
attach(dati)
head(dati) # visualizzo le prime righe del dataset
```

```
##       city year month sales volume median_price listings months_inventory
## 1 Beaumont 2010     1    83 14.162       163800     1533              9.5
## 2 Beaumont 2010     2   108 17.690       138200     1586             10.0
## 3 Beaumont 2010     3   182 28.701       122400     1689             10.6
## 4 Beaumont 2010     4   200 26.819       123200     1708             10.6
## 5 Beaumont 2010     5   202 28.833       123100     1771             10.9
## 6 Beaumont 2010     6   189 27.219       122800     1803             11.1
```

2. Indica il tipo di variabili contenute nel dataset.

variabile	descrizione	tipo
city	città	qualitativa e categoriale su scala nominale
year	anno di riferimento	qualitativa su scala ordinale
month	mese di riferimento	qualitativa su scala ordinale
sales	numero totale di vendite	quantitativa discreta su scala di rapporti
volume	valore totale delle vendite in milioni di dollari	quantitativa continua su scala di rapporti
median_price	prezzo mediano di vendita in dollari	quantitativa continua su scala di rapporti
listings	numero totale di annunci attivi	quantitativa discreta su scala di rapporti
months_inventory	quantità di tempo necessaria per vendere tutte le inserzioni correnti al ritmo attuale delle vendite, espresso in mesi.	quantitativa continua su scala di rapporti

3. Calcola Indici di posizione, variabilità e forma per tutte le variabili per le quali ha senso farlo, per le altre crea una distribuzione di frequenza. Commenta tutto brevemente.

Indici di POSIZIONE:

- moda
- media (aritmetica, ponderata, armonica e geometrica)
- mediana
- min e max
- percentili

Indici di VARIABILITA' E FORMA:

- Range o intervallo di variazione
- Differenza interquartile o range interquartile
- Varianza
- Deviazione standard o scarto quadratico medio
- Coefficiente di variazione
- Indice di eterogeneità di Gini (x variabili qualitative)

Indici di FORMA:

- asimmetria
- curtosi

var: **City**

Non è possibile calcolare i vari indici perchè è una variabile qualitativa, quindi calcolo la distribuzione di frequenza

```
ni = table(city) #freq assoluta (moda)
N = dim(dati)[1] #num righe e num colonne->prendo le righe
fi = table(city)/N # freq relativa
Ni = cumsum(ni) #freq cumulata
Fi = Ni/N #freq relative cumulate
# visualizziamo combinazione frequenze complete in tabella
table_complete = cbind(ni, fi, Ni, Fi)
table_complete
```

```
##                         ni   fi  Ni   Fi
## Beaumont                60 0.25  60 0.25
## Bryan-College Station 60 0.25 120 0.50
## Tyler                   60 0.25 180 0.75
## Wichita Falls           60 0.25 240 1.00
```

Non ha senso neanche calcolare gli indici per le variabili year e month che sono variabili qualitative su scala ordinale qunidi calcolo la distribuzione di frequenze.

var: **Year**

```
ni = table(year) #freq assoluta
N = dim(dati)[1] #num righe e num colonne->prendo le righe
fi = table(year)/N # freq relativa
Ni = cumsum(ni) #freq cumulata
Fi = Ni/N #freq relative cumulate
table_complete = cbind(ni, fi, Ni, Fi)
table_complete
```

```
##      ni  fi  Ni  Fi
## 2010 48 0.2  48 0.2
## 2011 48 0.2  96 0.4
## 2012 48 0.2 144 0.6
## 2013 48 0.2 192 0.8
## 2014 48 0.2 240 1.0
```

var: **Month**

```
ni = table(month) #freq assoluta
N = dim(dati)[1] #num righe e num colonne->prendo le righe
fi = table(month)/N # freq relativa
Ni = cumsum(ni) #freq cumulata
Fi = Ni/N #freq relative cumulate
# visualizziamo combinazione frequenze complete in tabella
table_complete = cbind(ni, fi, Ni, Fi)
table_complete
```

```
##    ni         fi  Ni         Fi
## 1  20 0.08333333  20 0.08333333
## 2  20 0.08333333  40 0.16666667
## 3  20 0.08333333  60 0.25000000
## 4  20 0.08333333  80 0.33333333
## 5  20 0.08333333 100 0.41666667
## 6  20 0.08333333 120 0.50000000
## 7  20 0.08333333 140 0.58333333
## 8  20 0.08333333 160 0.66666667
## 9  20 0.08333333 180 0.75000000
## 10 20 0.08333333 200 0.83333333
## 11 20 0.08333333 220 0.91666667
## 12 20 0.08333333 240 1.00000000
```

var: **Sales**

```
# indici di posizione
summary(sales)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    79.0   127.0   175.5   192.3   247.0   423.0
```

```r
# indici di variabilità
indici_variabilità_sales <- data.frame(
  R = max(sales) - min(sales),  # range o intervallo di variazione
  diff_interquartile = IQR(sales), #diff tra 3° e 1° quartile
  varianza = var(sales),
  dev_standard = sd(sales), #deviazione standard
  coeff_variazione = sd(sales)/mean(sales) * 100
)
print(indici_variabilità_sales)
```

```
##     R diff_interquartile varianza dev_standard coeff_variazione
## 1 344                120   6344.3     79.65111         41.42203
```

```r
# indici di forma
indici_forma_sales <- data.frame(
  asimmetria = skewness(sales),
  curtosi = kurtosis(sales) - 3
)
print(indici_forma_sales)
```

```
##   asimmetria    curtosi
## 1   0.718104 -0.3131764
```

var: **Volume**

```r
# indici di posizione
summary(volume)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   8.166  17.660  27.062  31.005  40.893  83.547
```

```r
# indici di variabilità
indici_variabilità_volume <- data.frame(
  R = max(volume) - min(volume),  # range o intervallo di variazione
  diff_interquartile = IQR(volume), #diff tra 3° e 1° quartile
  varianza = var(volume),
  dev_standard = sd(volume), #deviazione standard
  coeff_variazione = sd(volume)/mean(volume) * 100
)
print(indici_variabilità_volume)
```

```
##        R diff_interquartile varianza dev_standard coeff_variazione
## 1 75.381            23.2335 277.2707     16.65145         53.70536
```

```r
# indici di forma
indici_forma_volume <- data.frame(
  asimmetria = skewness(volume),
  curtosi = kurtosis(volume) - 3
)
print(indici_forma_volume)
```

```
##   asimmetria  curtosi
## 1   0.884742 0.176987
```

var: **Median_price**

```r
# indici di posizione
summary(median_price)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   73800  117300  134500  132665  150050  180000
```

```r
# indici di variabilità
indici_variabilità_price <- data.frame(
  R = max(median_price) - min(median_price),  # range o intervallo di variazione
  diff_interquartile = IQR(median_price), #diff tra 3° e 1° quartile
  varianza = var(median_price),
  dev_standard = sd(median_price), #deviazione standard
  coeff_variazione = sd(median_price)/mean(median_price) * 100
)
print(indici_variabilità_price)
```

```
##        R diff_interquartile  varianza dev_standard coeff_variazione
## 1 106200              32750 513572983     22662.15         17.08218
```

```r
# indici di forma
indici_forma_median_price <- data.frame(
  asimmetria = skewness(median_price),
  curtosi = kurtosis(median_price) - 3
)
print(indici_forma_median_price)
```

```
##   asimmetria    curtosi
## 1 -0.3645529 -0.6229618
```

var: **Listings**

```r
# indici di posizione
summary(listings)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     743    1026    1618    1738    2056    3296
```

```r
# indici di variabilità
indici_variabilità_listings <- data.frame(
  R = max(listings) - min(listings),  # range o intervallo di variazione
  diff_interquartile = IQR(listings), #diff tra 3° e 1° quartile
  varianza = var(listings),
  dev_standard = sd(listings), #deviazione standard
  coeff_variazione = sd(listings)/mean(listings) * 100
)
print(indici_variabilità_listings)
```

```
##      R diff_interquartile varianza dev_standard coeff_variazione
## 1 2553             1029.5   566569     752.7078         43.30833
```

```r
# indici di forma
indici_forma_listings <- data.frame(
  asimmetria = skewness(listings),
  curtosi = kurtosis(listings) - 3
)
print(indici_forma_listings)
```

```
##   asimmetria  curtosi
## 1  0.6494982 -0.79179
```

var: **months_inventory**

```r
# indici di posizione
summary(months_inventory)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.400   7.800   8.950   9.193  10.950  14.900
```

```r
# indici di variabilità
indici_variabilità_inventory <- data.frame(
  R = max(months_inventory) - min(months_inventory),  # range o intervallo di variazione
  diff_interquartile = IQR(months_inventory), #diff tra 3° e 1° quartile
  varianza = var(months_inventory),
  dev_standard = sd(months_inventory), #deviazione standard
  coeff_variazione = sd(months_inventory)/mean(months_inventory) * 100
)
print(indici_variabilità_inventory)
```

```
##      R diff_interquartile varianza dev_standard coeff_variazione
## 1 11.5               3.15 5.306889     2.303669         25.06031
```

```r
# indici di forma
indici_forma_months_inventory <- data.frame(
  asimmetria = skewness(months_inventory),
  curtosi = kurtosis(months_inventory) - 3
)
print(indici_forma_months_inventory)
```

```
##   asimmetria    curtosi
## 1 0.04097527 -0.1744475
```

4. Qual è la variabile con variabilità più elevata? Come ci sei arrivato? E quale quella più asimmetrica?

- Per confrontare la variabilità di un campione tra due diverse variabili, si utilizza il **coefficente di variazione**. Dai nostri dati emerge che la variabile **volume** presenta una maggiore variabilità relativa rispetto alla media, con un coefficiente di variazione del 53.71% circa.
- **volume** è anche la variabile più asimmetrica perchè basandomi sull'indice di asimmetria di Fisher, ha il valore che più si allontana dallo 0 (quindi da una distribuzione simmetrica)

5. Dividi una delle variabili quantitative in classi, scegli tu quale e come, costruisci la distribuzione di frequenze, il grafico a barre corrispondente e infine calcola l'indice di Gini.

Suddivisione in classi di variabile quantitativa (basandosi su lunghezza)

```r
min(sales) #prendo il minimo
```

```
## [1] 79
```

```r
max(sales) #prendo il massimo
```

```
## [1] 423
```

```r
# creo nuova colonna
class = seq(50, 450, by=50)
class_sales = cut(sales, breaks = class)
```

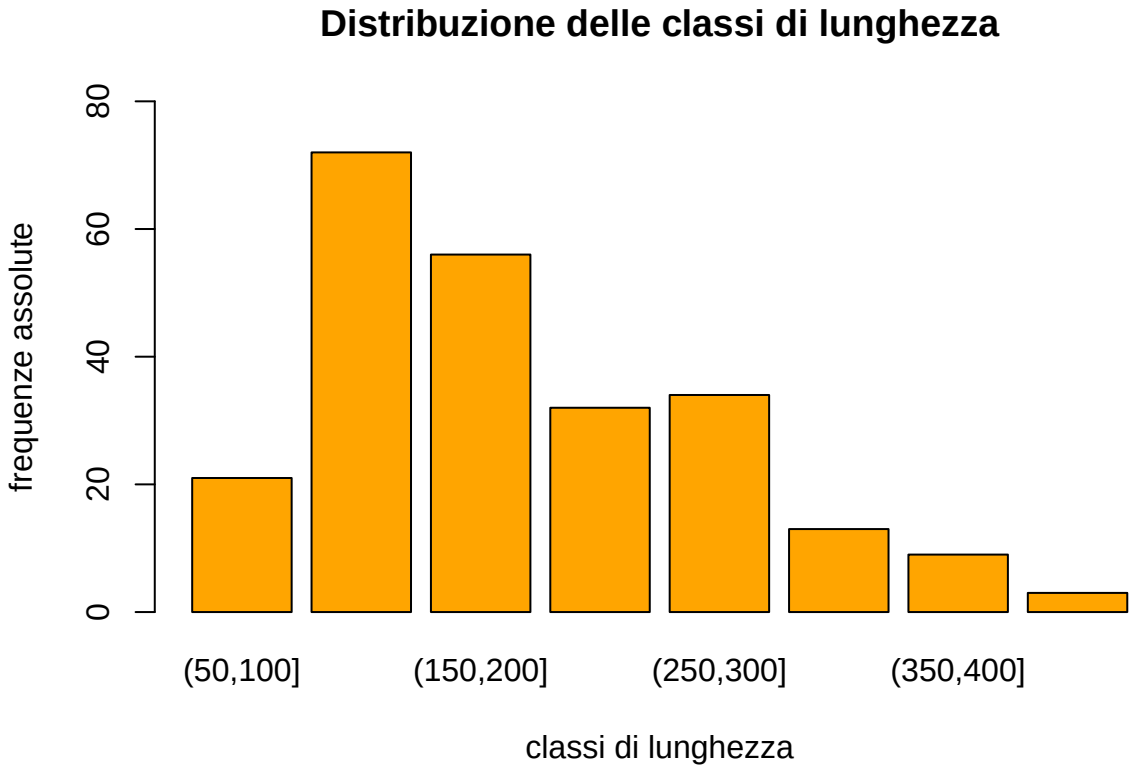
Distribuzione di frequenze

```r
ni = table(class_sales) #freq assoluta
N = dim(dati)[1] #num righe e num colonne->prendo le righe
fi = table(class_sales)/N # freq relativa
Ni = cumsum(ni) #freq cumulata
Fi = Ni/N #freQ relative cumulate
# visualizziamo combinazione frequenze complete in tabella
table_complete = cbind(ni, fi, Ni, Fi)
```

Grafico a barre corrispondente

```r
distr_freq = as.data.frame(table_complete)

?barplot
barplot(ni,
        main = "Distribuzione delle classi di lunghezza",
        xlab = "classi di lunghezza",
        ylab = "frequenze assolute",
        ylim = c(0,80),
        col = "orange")
```



Indice di Gini

```r
gini.index <- function(x){
  ni = table(x)
  fi = ni/length(x)
  fi2 = fi^2
  J = length(table(x))

  gini = 1-sum(fi2)
  gini.norm = gini/((J-1)/J)
```

```
  return(gini.norm)
}
table(class_sales)
```

```
## class_sales
##  (50,100] (100,150] (150,200] (200,250] (250,300] (300,350] (350,400] (400,450]
##        21        72        56        32        34        13         9         3
```

```
gini.index(class_sales)
```

```
## [1] 0.9206349
```

L'indice di Gini ottenuto è elevato, questo si traduce in un alto livello di diseguaglianza tra le varie classi.

7. Qual è la probabilità che presa una riga a caso di questo dataset essa riporti la città "Beaumont"? E la probabilità che riporti il mese di Luglio? E la probabilità che riporti il mese di dicembre 2012?

$$P = \frac{n\check{r}casiFavorevoli}{n\check{r}casiTotali}$$

Probabilità che presa una riga a caso dal dataset riporti la città "Beaumont"

```
num_osservazioni <- nrow(dati) #tot righe dataset
num_city_beaumont <- nrow(subset(dati, city == "Beaumont")) #tot righe con città = Beaumont
p_city_Beaumont = num_city_beaumont/num_osservazioni
p_city_Beaumont
```

```
## [1] 0.25
```

Probabilità che presa una riga a caso dal dataset riporti il mese di "Luglio"

```
num_month_luglio <- nrow(subset(dati, month == "7"))
p_month_july = num_month_luglio/num_osservazioni
p_month_july
```

```
## [1] 0.08333333
```

Probabilità che presa una riga a caso dal dataset riporti il mese di "Dicembre" e l'anno "2012"

```
num_dicembre_2012 <- nrow(subset(dati, month == "12" & year == 2012))
p_dicembre_2012 <- num_dicembre_2012 / num_osservazioni
p_dicembre_2012
```

```
## [1] 0.01666667
```

8. Esiste una colonna col prezzo mediano, creane una che indica invece il prezzo medio, utilizzando le altre variabili che hai a disposizione.

```
mutate(
  dati,
  mean_price = volume / sales) %>% select(volume, sales, mean_price) %>% head(5)
```

```
##   volume sales mean_price
## 1 14.162    83  0.1706265
## 2 17.690   108  0.1637963
## 3 28.701   182  0.1576978
## 4 26.819   200  0.1340950
## 5 28.833   202  0.1427376
```

9. Prova a creare un'altra colonna che dia un'idea di "efficacia" degli annunci di vendita. Riesci a fare qualche considerazione?

```
mutate(
  dati,
  efficiency_listings = sales / listings)%>%
select(sales, listings, efficiency_listings) %>% head(5)
```

```
##   sales listings efficiency_listings
## 1    83     1533          0.05414220
## 2   108     1586          0.06809584
## 3   182     1689          0.10775607
## 4   200     1708          0.11709602
## 5   202     1771          0.11405985
```

Considerazione:

Esempi:

- 83 vendite e 1533 annunci -> efficiacia annuncio 0.05
- 200 vednite e 1708 annunci -> efficacia annuncio 0.11
- 124 vendite e 1829 annunci -> efficacia annuncio 0.06
- 83 vendite e 1533 annunci -> efficacia annuncio 0.05

In media, sembra che il numero di annunci utilizzati per effettuare le vendite non influenzi in modo significativo l'efficacia delle vendite, poiché l'efficacia rimane relativamente costante nei diversi scenari presentati.

10. Prova a creare dei summary(), o semplicemente media e deviazione standard, di alcune variabili a tua scelta, condizionatamente alla città, agli anni e ai mesi. Puoi utilizzare il linguaggio R di base oppure essere un vero Pro con il pacchetto dplyr. Ti lascio un suggerimento in pseudocodice, oltre al cheatsheet nel materiale:

```
# summary by year
dati %>%
  group_by(year) %>%
  summarise(
    mean = mean(listings, na.rm = TRUE),
    sd = sd(listings, na.rm = TRUE),
    min = min(listings),
    max = max(listings))
```

```
## # A tibble: 5 x 5
##    year  mean    sd   min   max
##   <int> <dbl> <dbl> <int> <int>
## 1  2010 1826   785.   904  3296
## 2  2011 1850.  780.   844  3266
## 3  2012 1777.  738.   801  3072
## 4  2013 1678.  744.   743  2998
## 5  2014 1560.  707.   746  2875
```

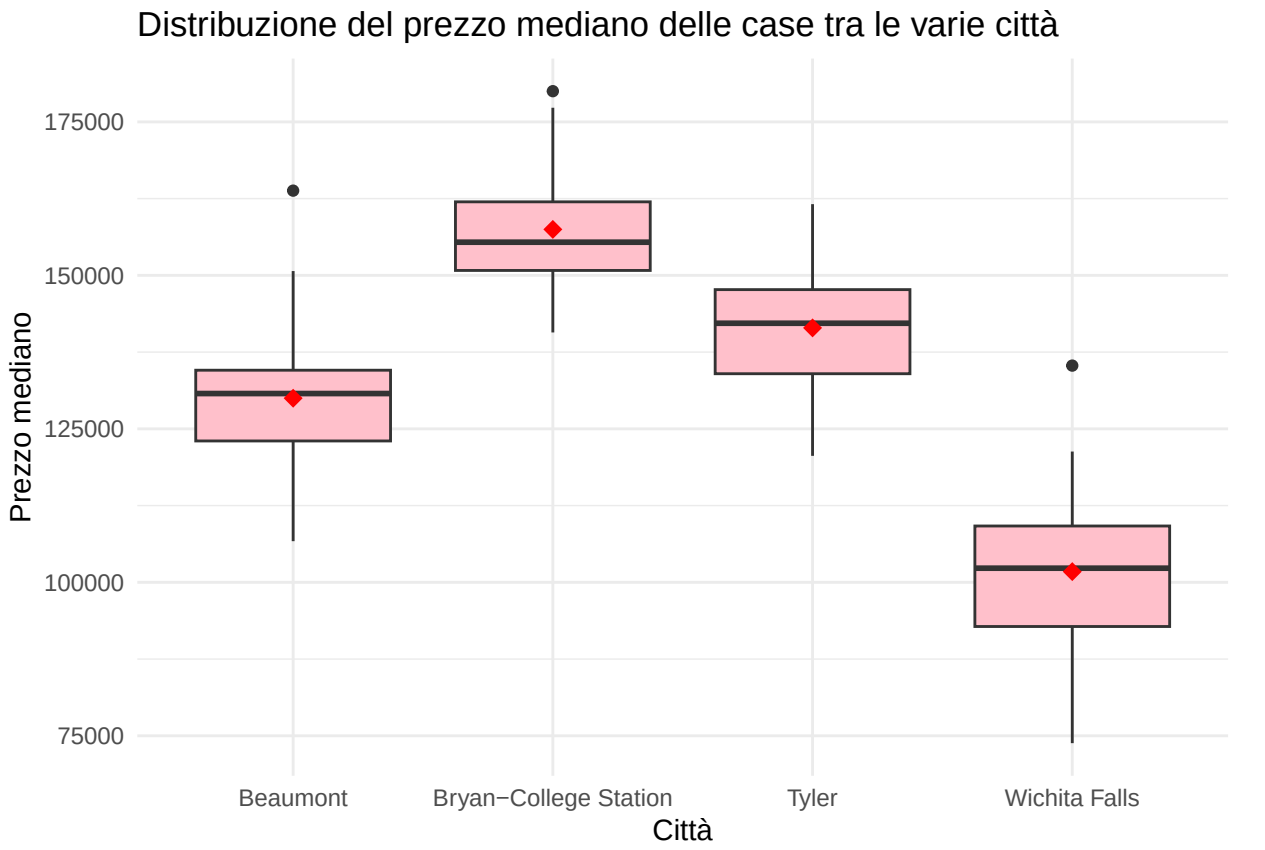
```
# summary by city, year and month
dati %>%
  group_by(city, year, month) %>%
  summarise(
    media_sales = mean(sales, na.rm = TRUE),  .groups = "rowwise") %>% head(5)
```

```
## # A tibble: 5 x 4
## # Rowwise:  city, year, month
##   city      year month media_sales
##   <chr>    <int> <int>       <dbl>
## 1 Beaumont  2010     1          83
## 2 Beaumont  2010     2         108
## 3 Beaumont  2010     3         182
## 4 Beaumont  2010     4         200
## 5 Beaumont  2010     5         202
```

GRAFICI

1. Utilizza i boxplot per confrontare la distribuzione del prezzo mediano delle case tra le varie città. Commenta il risultato

```
ggplot(dati, aes(x = city, y = median_price)) +
  geom_boxplot(fill = "pink", ) +
   stat_summary(fun = "mean", geom = "point", shape = 18, size = 3, color = "red") +
  labs(title = "Distribuzione del prezzo mediano delle case tra le varie città",
       x = "Città",
       y = "Prezzo mediano")+
  theme_minimal()
```



Considerazioni:

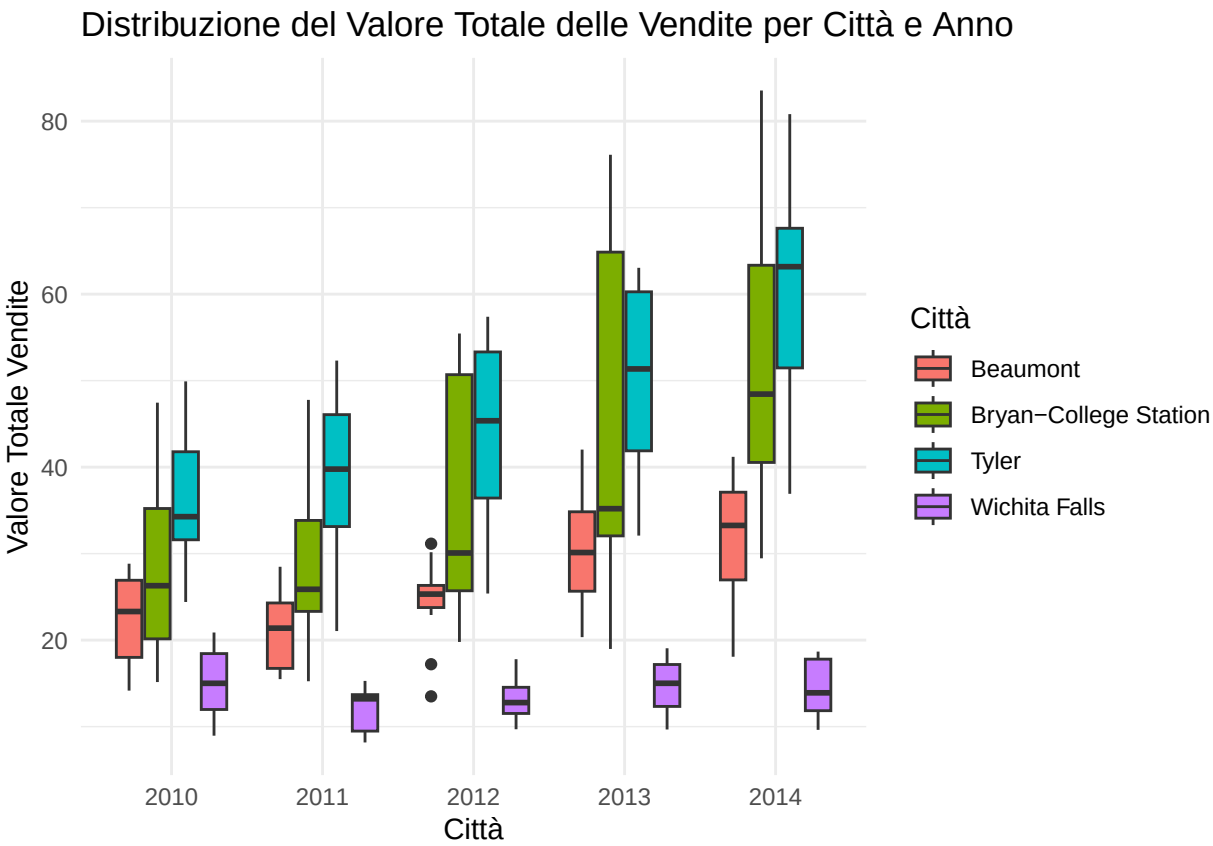
- Centralità della distribuzione: Si tratta di una distribuzione asimmetrica perchè la mediana non è al centro (ma è spostata in alcuni casi più verso l'alto e in altri più verso il basso)
- Distribuzione: ci sono variazioni nella distribuzione dei prezzi mediani tra le città in quanto i boxplot

non sono sulla stessa riga (hanno prezzo mediano molto diverso)

- Variabilità dei dati: bassa dispersione dei dati in quanto la lunghezza dei box e anche le code sono approssimitivamente le stesse
- Presenza di valori anomali: Nelle città di "Beaumont", "Bryan-College Station", e "Wichita Falls" sono presenti dei valori anomali (punti esterni alle barre)

2. Utilizza i boxplot o qualche variante per confrontare la distribuzione del valore totale delle vendite tra le varie città ma anche tra i vari anni. Qualche considerazione da fare?

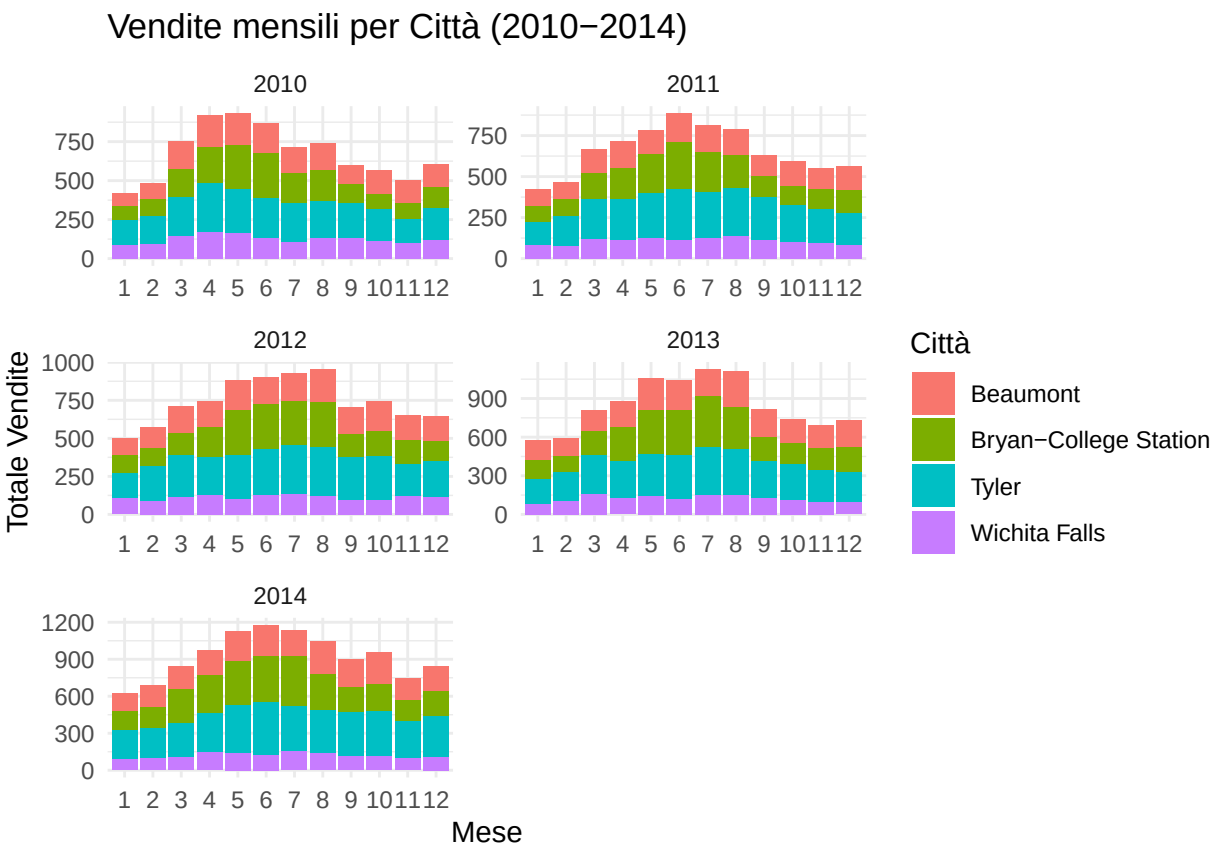
```
ggplot(dati, aes(x = factor(year), y = volume, fill = city)) +
  geom_boxplot() +
  labs(title = "Distribuzione del Valore Totale delle Vendite per Città e Anno",
       x = "Città",
       y = "Valore Totale Vendite",
       fill = "Città") +
  theme_minimal()
```



Nel periodo compreso tra il 2012 e il 2014, le città di Bryan-College-Station e Tyler hanno evidenziato prestazioni di vendita superiori rispetto a Beaumont e Wichita Falls. Questo trend indica un notevole successo commerciale per le prime due città durante quel triennio specifico.

3. Usa un grafico a barre sovrapposte per confrontare il totale delle vendite nei vari mesi, sempre considerando le città. Prova a commentare ciò che viene fuori. Già che ci sei prova anche il grafico a barre normalizzato. Consiglio: Stai attento alla differenza tra geom_bar() e geom_col(). PRO LEVEL: cerca un modo intelligente per inserire ANCHE la variabile Year allo stesso blocco di codice, senza però creare accrocchi nel grafico.

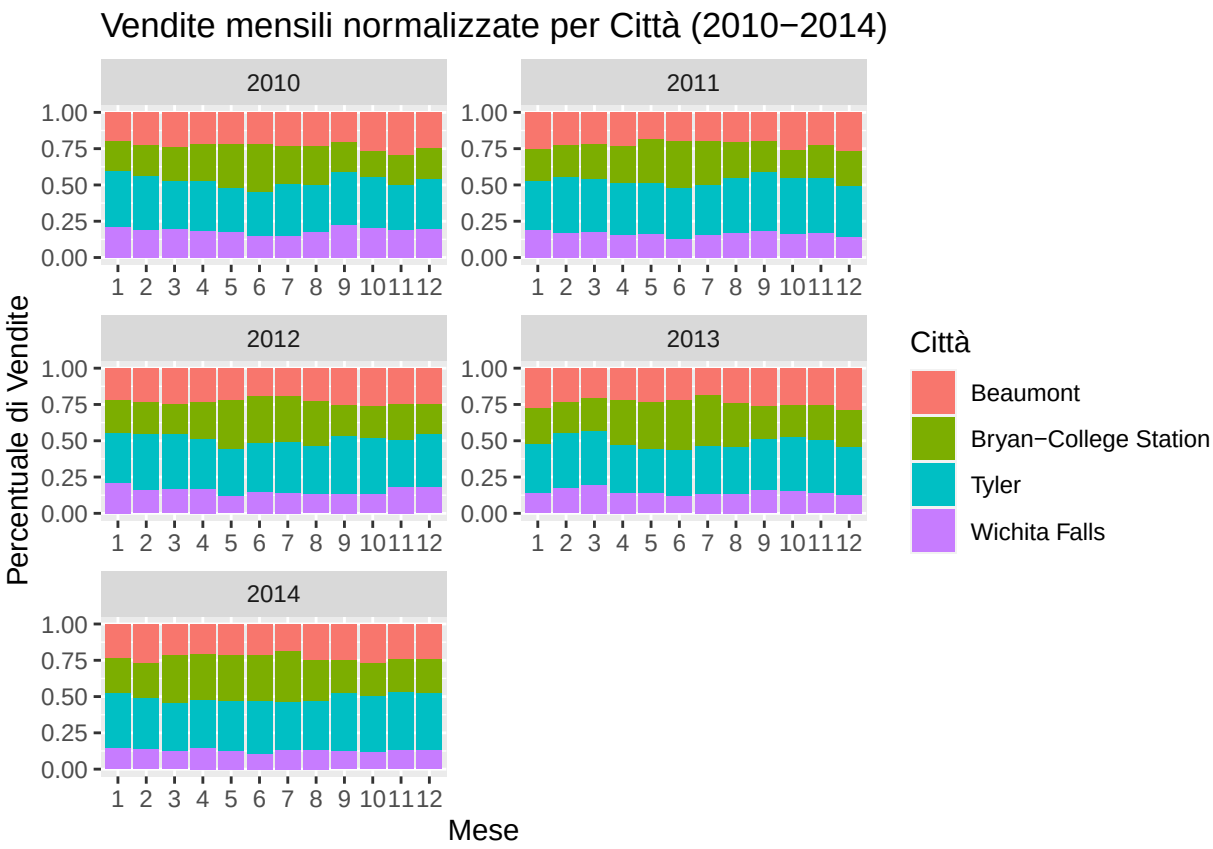
```
# Grafico a barre sovrapposte
ggplot(dati, aes(x = factor(month), y = sales, fill = city)) +
  geom_bar(stat = "identity", position = "stack") + #stack per barre sovrapposte
  facet_wrap(~year, scales = "free", ncol = 2) +
  labs(title = "Vendite mensili per Città (2010-2014)",
       x = "Mese",
       y = "Totale Vendite",
       fill = "Città") +
  theme_minimal()
```



Per ciascun anno nel periodo dal 2010 al 2014, si è osservato un incremento significativo nelle vendite durante i mesi estivi, vale a dire maggio, giugno, luglio e agosto. Questa tendenza è particolarmente evidente nella città di Bryan-College Station, dove le vendite risultano essere consistentemente superiori in questi mesi specifici.

```
# Grafico a barre normalizzato
ggplot(dati, aes(x = factor(month), y = sales, fill = city)) +
  geom_bar(stat = "identity", position = "fill") +  # "fill" per normalizzare le barre
  facet_wrap(~year, scales = "free", ncol = 2) +
  labs(title = "Vendite mensili normalizzate per Città (2010-2014)",
       x = "Mese",
       y = "Percentuale di Vendite",
```

```
        fill = "Città")
```



4. Prova a creare un line chart di una variabile a tua scelta per fare confronti commentati fra città e periodi storici. Ti avviso che probabilmente all'inizio ti verranno fuori linee storte e poco chiare, ma non demordere. Consigli: Prova inserendo una variabile per volta. Prova a usare variabili esterne al dataset, tipo vettori creati da te appositamente.

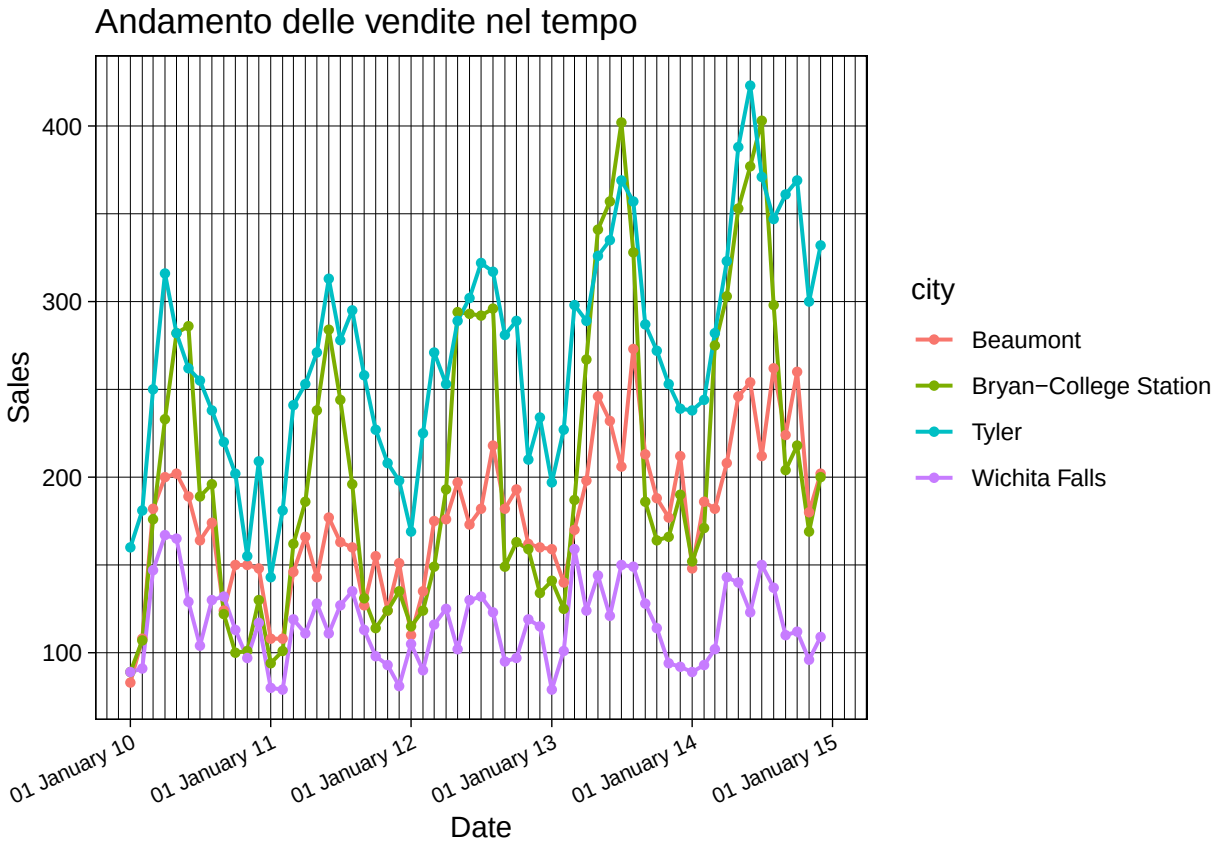
```
year_month <- as.Date(paste(dati$year, dati$month, "01", sep = "-"))
length(year_month)
```

```
## [1] 240
```

```
length(dati$sales)
```

```
## [1] 240
```

```
ggplot(dati, aes(x = year_month, y = sales, group=city, color=city)) +
  geom_line(linewidth = 0.7) +
  geom_point(size = 1.5, shape = 16) +  # Aggiungi i punti
  labs(title = "Andamento delle vendite nel tempo",
       x = "Date",
       y = "Sales") +
  scale_x_date(labels = scales::date_format("%d %B %y"), date_breaks = "1 year", date_minor_breaks = "1
  theme_linedraw()+
  theme(axis.text.x = element_text(angle = 25, hjust = 1, vjust = 1, size = 8))
```



Considerazioni: Sto esaminando il confronto tra il numero totale di vendite (sales) registrate ogni mese nel periodo compreso tra il 2010 e il 2014. In questo contesto, emerge chiaramente una tendenza di vendite superiori nelle città di Bryan-College Station e Tyler nel corso di questi anni, con un picco più elevarto intorno a maggio 2014 per la città di Tyler.