# Final_project

Elisa Lipari

2024-01-12

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

```
setwd("C:/Users/SAMSUNG/Desktop")

getwd()

## [1] "C:/Users/SAMSUNG/Desktop"

data=read.csv("realestate_texas.csv")


#visualizzo le prime 5 righe del dataset:

head(data,5)

##       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

#valuta la dimensione del mio datset
dim (data)

## [1] 240   8

attach(data)
```

2) Indica il tipo di variabili del dataset

• City: variabile qualitativa su scala nominale

• Year: variabile quantitativa continua

• Month: variabile qualitativa ordinata ciclica

• Sales: variabile quantitativa discreta

• Volume: variabile quantitativa continua

• Median_price: variabile quantitativa continua

• listings: variabile quantitativa discreta

• months_inventory: variabile quantitativa continua

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.

- CITY

Indici di posizione: moda

```
#moda
table(city)

## city
##               Beaumont Bryan-College Station                 Tyler
##                     60                     60                    60
##         Wichita Falls
##                     60
```

*city ha una distribuzione quadrimodale.*

- YEAR

Indici di posizione: moda

Indice di variabilità: range

```
#calcolo range

range(year)

## [1] 2010 2014

#calcolo moda

table(year)

## year
## 2010 2011 2012 2013 2014
##   48   48   48   48   48
```

- MONTH

calcolo distribuzione di frequenze

Variabilità: range e gini

```
#calcolo distribuzione di frequenze:

N=dim(data)[1]
freq_assoluta=table(data["month"])
freq_relativa= table(data["month"])/N
```

```r
distr_freq=cbind(freq_assoluta,freq_relativa)
distr_freq
```

```
##    freq_assoluta freq_relativa
## 1             20    0.08333333
## 2             20    0.08333333
## 3             20    0.08333333
## 4             20    0.08333333
## 5             20    0.08333333
## 6             20    0.08333333
## 7             20    0.08333333
## 8             20    0.08333333
## 9             20    0.08333333
## 10            20    0.08333333
## 11            20    0.08333333
## 12            20    0.08333333
```

```r
#calcolo range

range(month)
```

```
## [1]  1 12
```

```r
#calcolo gini

calculate_gini <- function(x) {
  ni=table(x)
  fi=ni/length (x)
  fi2= fi^2
  J=length(table(x))
  gini=1-sum (fi2)
  gini.normalizzato=gini/((J-1)/J)
  return (gini.normalizzato)}

calculate_gini (month)
```

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

*la varibile month è equidistribuita ed ogni mese si presenta 20 volte nel dataset*

- SALES

Indici di posizione: Media, mediana, quartili

Variabilità: range, Deviazione standard, varianza, differenza interquartile.

Forma: Skewness, kurtosis.

```r
#media
mean (sales)
```

```
## [1] 192.2917
```

```r
#mediana
median(sales)
```

```
## [1] 175.5
```

```r
#quartili

quantile(sales)
```

```
##    0%   25%   50%   75%  100%
##  79.0 127.0 175.5 247.0 423.0
```

```r
#range
range(sales)
```

```
## [1]  79 423
```

```r
#varianza, deviazione std, differenza interquartile
sd(sales)
```

```
## [1] 79.65111
```

```r
var(sales)
```

```
## [1] 6344.3
```

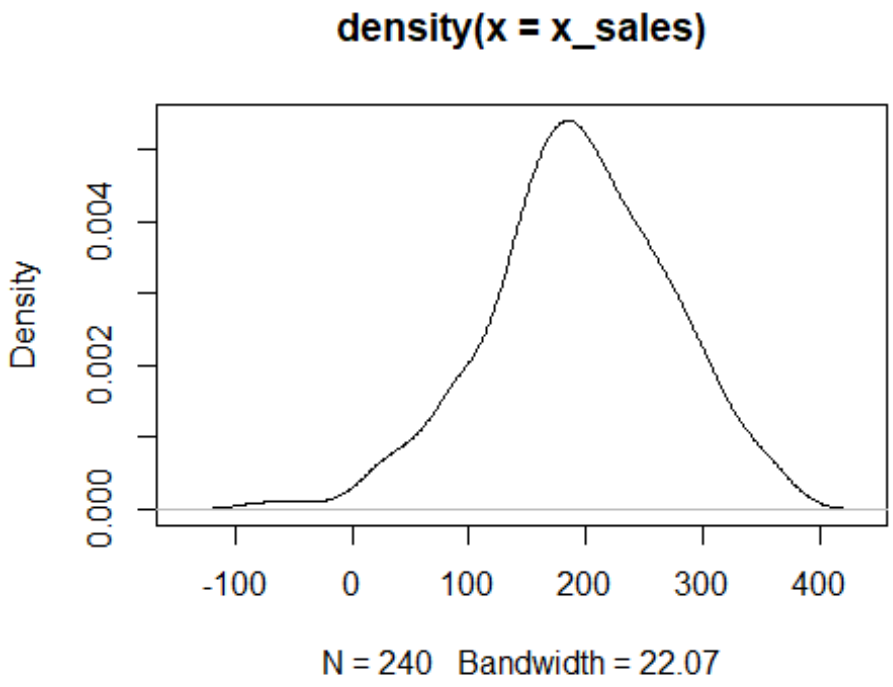
```r
IQR(sales)
```

```
## [1] 120
```

```r
#CV

cv=function(x){return (sd(x)/mean(x)*100)}
cv(sales)
```

```
## [1] 41.42203
```

```r
#Skewness, kurtosis

install.packages("moments")

library(moments)

x_sales= rnorm(sales,mean(sales), sd(sales))
plot(density(x_sales))
```




```
skewness(x_sales)
```

```
## [1] -0.2905999
```

```
kurtosis(x_sales)-3
```

```
## [1] 0.2575336
```

*la variabile Sales ha una distribuzione asimmetrica positiva e leptocurtica*

- VOLUME

Indici di posizione: tramite funzione summary()

Variabilità: Deviazione standard, varianza, range interquartile e CV.

Forma: Skewness, kurtosis.

```
#funzione summary():
```

```
summary(volume)
```

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

```
#deviazione standard
```

```
sd(volume)
```

```
## [1] 16.65145
```

```r
#varianza

var(volume)
```

```
## [1] 277.2707
```

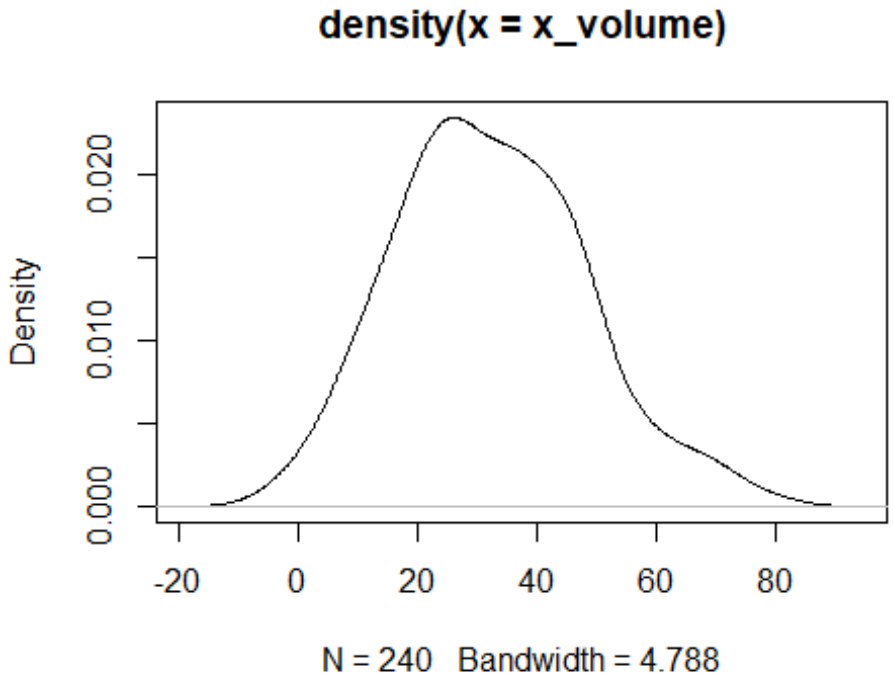
```r
#range interquartile

IQR(volume)
```

```
## [1] 23.2335
```

```r
#CV

cv(volume)
```

```
## [1] 53.70536
```

```r
#indici di forma:

x_volume= rnorm(volume,mean(volume), sd(volume))
plot(density(x_volume))
```




```r
skewness(x_volume)
```

```
## [1] 0.3220528
```

```r
kurtosis(x_volume)-3
```

```
## [1] -0.0760788
```

*la variabile Volume ha una distribuzione asimmetrica negativa e leptocurtica*

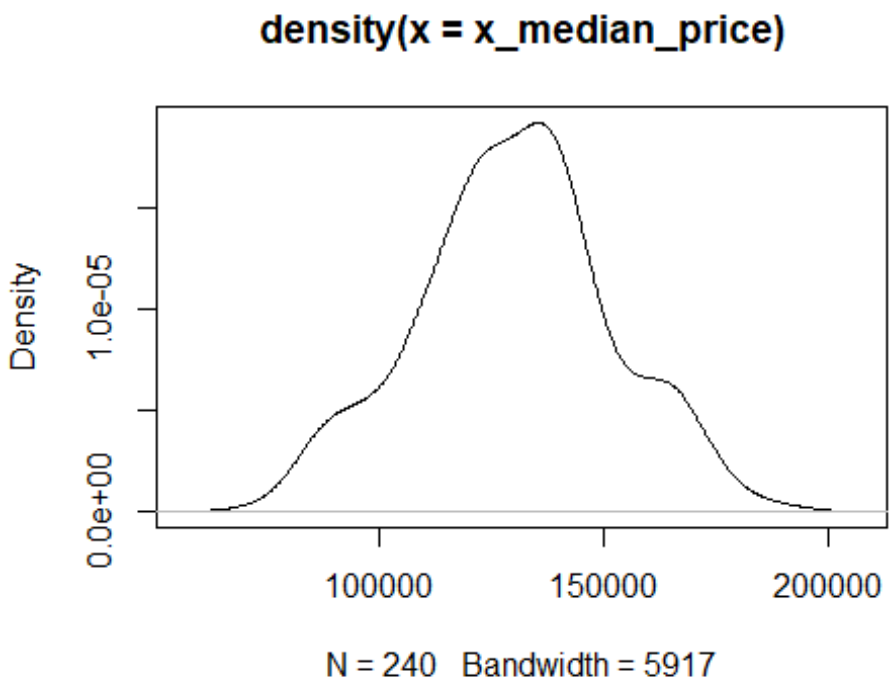
- MEDIAN_PRICE

Indici di posizione: tramite funzione summary()

Variabilità: Deviazione standard, varianza, range interquartile e CV.

Forma: Skewness, kurtosis.

```
#funzione summary():

summary(median_price)

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

#deviazione standard

sd(median_price)

## [1] 22662.15

#varianza

var(median_price)

## [1] 513572983

#range interquartile

IQR(median_price)

## [1] 32750

#CV

cv(median_price)

## [1] 17.08218

#indici di forma:

x_median_price= rnorm(median_price,mean(median_price), sd(median_price))
plot(density(x_median_price))
```




```
skewness(x_median_price)
```

```
## [1] -0.01292076
```

```
kurtosis(x_median_price)-3
```

```
## [1] -0.176612
```

*la varibile median_price ha una distribuzione asimmetrica positiva e platicurtica*

- LISTINGS

Indici di posizione: tramite funzione summary()

Variabilità: Deviazione standard, varianza, range interquartile e CV.

Forma: Skewness, kurtosis.

```
#funzione summary():

summary(listings)
```

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

```
#deviazione standard

sd(listings)
```

```
## [1] 752.7078
```

```
#varianza
```

```
var(listings)
```

```
## [1] 566569
```

```
#range interquartile
```

```
IQR(listings)
```

```
## [1] 1029.5
```

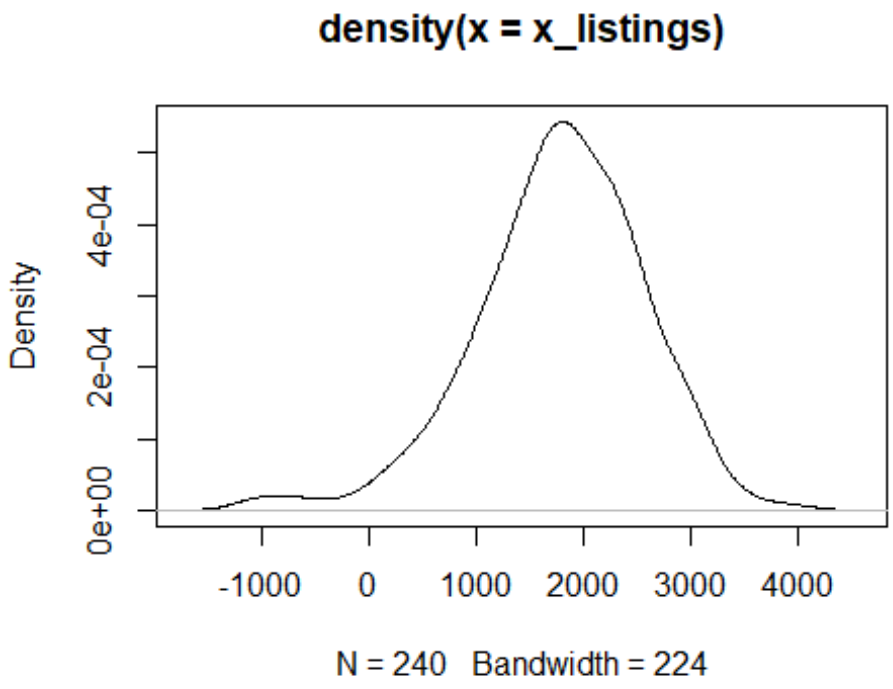
```
#CV
```

```
cv(listings)
```

```
## [1] 43.30833
```

```
#indici di forma:
```

```
x_listings= rnorm(listings,mean(listings), sd(listings))
plot(density(x_listings))
```




```
skewness(x_listings)
```

```
## [1] -0.5941654
```

```
kurtosis(x_listings)-3
```

```
## [1] 1.09179
```

*la varibile Listings ha una distribuzione leggermente asimmetrica positiva e leptocurtica*

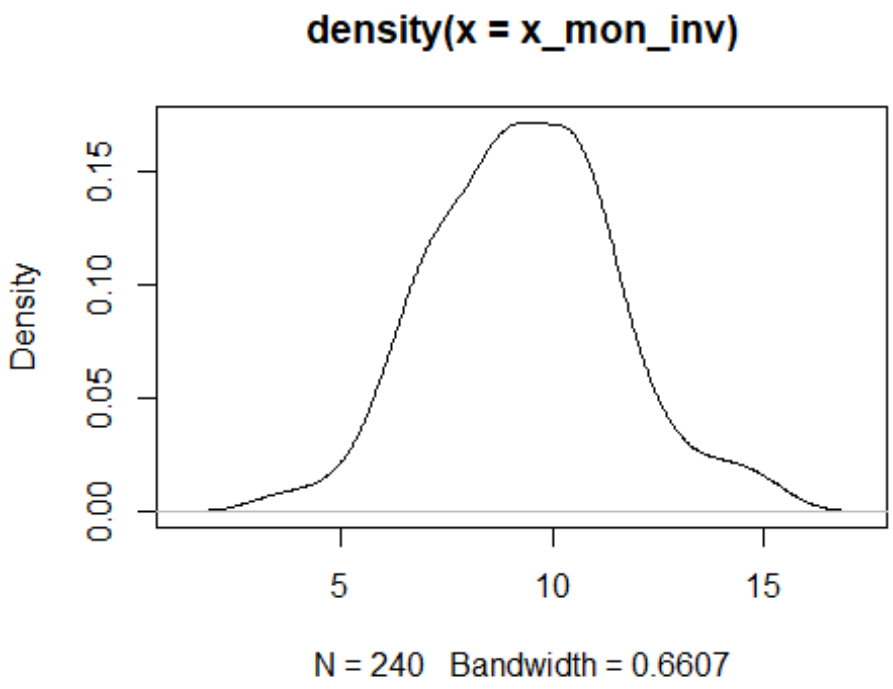
- MONTHS_INVENTORY

Indici di posizione: tramite funzione summary()

Variabilità: Deviazione standard, varianza, range interquartile e CV.

Forma: Skewness, kurtosis.

```
#funzione summary():

summary(months_inventory)

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

#deviazione standard

sd(months_inventory)

## [1] 2.303669

#varianza

var(months_inventory)

## [1] 5.306889

#range interquartile

IQR(months_inventory)

## [1] 3.15

#CV

cv(months_inventory)

## [1] 25.06031

#indici di forma:

x_mon_inv= rnorm(months_inventory,mean(months_inventory), sd(months_inventory))
plot(density(x_mon_inv))
```




```
skewness(x_mon_inv)

## [1] 0.1340054

kurtosis(x_mon_inv)-3

## [1] 0.1410895
```

*la varibile Month_inventory ha una distribuzione asimmetrica positiva e platicurtica*

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

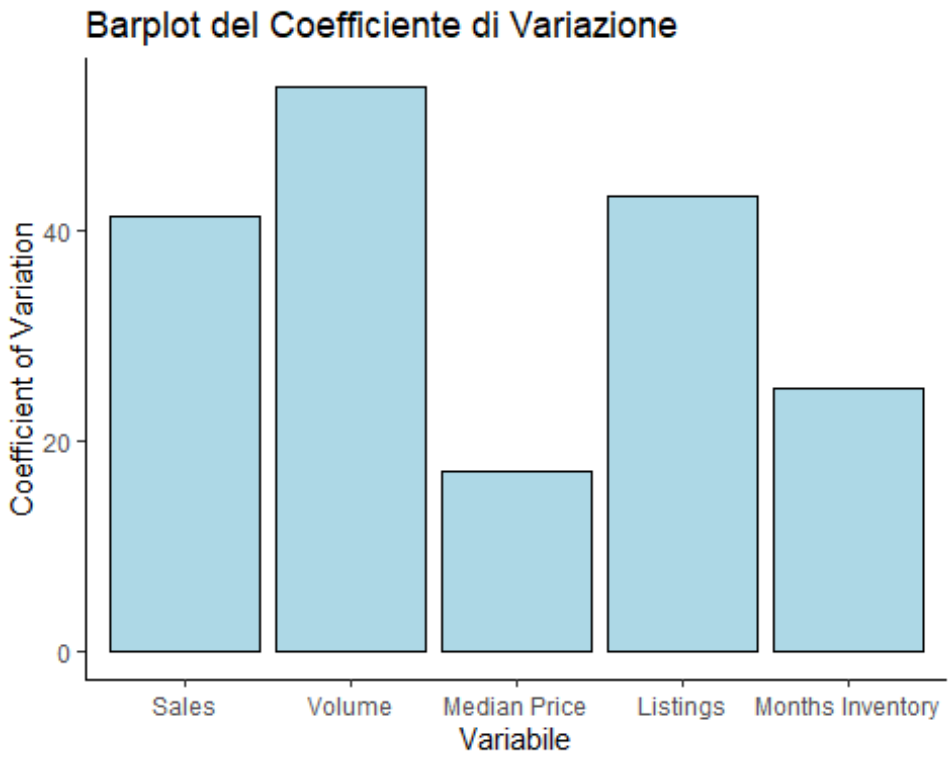
*confronto i CV delle variabili quantitative per valutare la variabilità maggiore:*

```
install.packages("ggplot2")

library(ggplot2)

all_cv= c(cv(sales), cv(volume), cv(median_price), cv(listings), cv(months_invento
ry))

all_cv

## [1] 41.42203 53.70536 17.08218 43.30833 25.06031
```

*Volume ha la variabilità più alta.*

*Per osservarlo graficamente:*

```
# Crea un data frame con nome variabili e valori cv

dati= data.frame(variable=c("Sales", "Volume", "Median Price", "Listings", "Months
Inventory"),
cv_value = all_cv)

# per avere in grafico l'ordine da me scelto
dati$variable= factor(dati$variable, levels = dati$variable)

# Creo il grafico per CV values

library(ggplot2)

grafico=ggplot(dati) +
  geom_bar(aes(x = variable, y = cv_value),
           stat = "identity",
           col="black",
           fill="lightblue") +
  labs(title = "Barplot del Coefficiente di Variazione",
       x = "Variabile",
       y = "Coefficient of Variation")+
  theme_classic()

grafico
```




*la variabile volume ha un coefficiente di variazione più alto rispetto alle altre variabili*

*Per valutare la variabile più asimmetrica:*

```
#per calcolo asimmetria
library(moments)

# creo un vettore con skewness
skewness = c(skewness(sales), skewness(volume), skewness(median_price), skewness(l
istings), skewness(months_inventory))

#mando la funzione per valutare la variabile con asimmetria maggiore

skewness
```

```
## [1]  0.71810402  0.88474203 -0.36455288  0.64949823  0.04097527
```

*la variabile volume ha una asimmetria maggiore rispetto alle altre variabili*

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.

*scelgo la variabile sales e la divido in classe:*

```
min(sales)
```

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

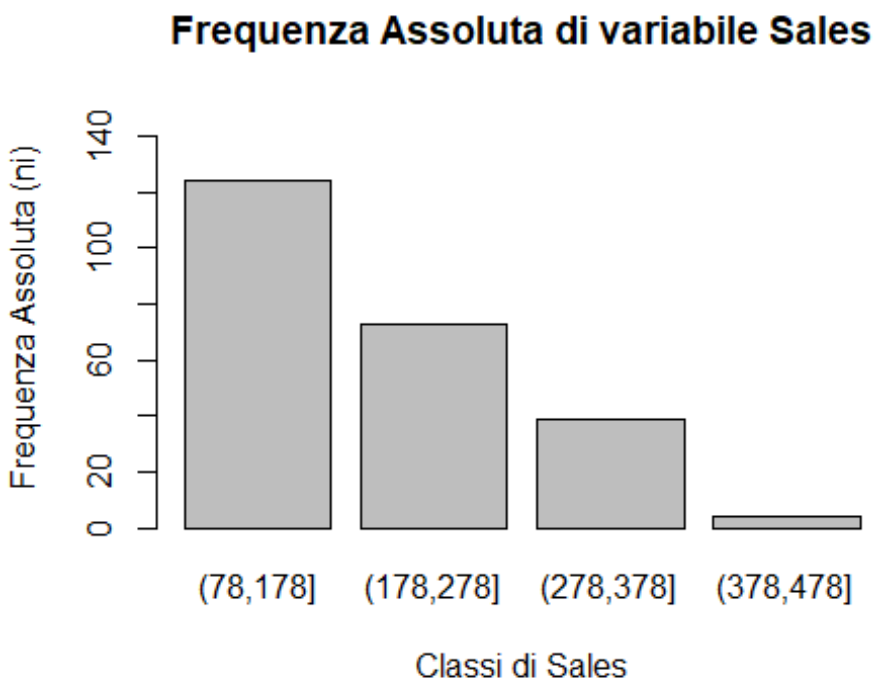
```
max(sales)
```

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

```
#creo la classe sales_cl con la funzione cut

sales_cl= cut (sales, breaks = c(78, 178, 278,378,478))


#creo la distribuzione di frequenze di sales

ni=table (sales_cl)
fi=table(sales_cl)/N
Ni=cumsum(ni)
Fi=Ni/N

#la visualizzo in forma tabellare:

distr_freq=as.data.frame (cbind(ni,fi,Ni,Fi))
distr_freq
```

```
##             ni         fi  Ni        Fi
## (78,178]  124 0.51666667 124 0.5166667
## (178,278]  73 0.30416667 197 0.8208333
## (278,378]  39 0.16250000 236 0.9833333
## (378,478]   4 0.01666667 240 1.0000000
```

```r
#faccio il grafico delle frequenze assolute

Freq=barplot(distr_freq$ni,
        main = "Frequenza Assoluta di variabile Sales",
        xlab = "Classi di Sales",
        ylab = "Frequenza Assoluta (ni)",
        ylim = c(0, 150),
        names.arg = rownames(distr_freq))
```




*La classe che viene osservata con maggior frequenza è quell ache contiene valori che vanno da 79 a 178.*

```r
#calcolo indice di gini

calculate_gini=function(x) {
  ni=table(x)
  fi=ni/length (x)
  fi2= fi^2
  J=length(table(x))
  gini=1-sum (fi2)
  gini.normalizzato=gini/((J-1)/J)
  return (gini.normalizzato)
  }


calculate_gini (sales_cl)
```

```
## [1] 0.8184722
```

*L' indice di Gini per sales_cl indica una distribuzione abbastanza disuguale tra le classi*

6) Indovina l'indice di gini per la variabile city.

```
calculate_gini(city)
```

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

*l'indice di gini è 1 ed indica massima eterogeneità.*

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?

```
#probabilità città beaumont

table(city)
```

```
## city
##              Beaumont Bryan-College Station                 Tyler
##                    60                    60                    60
##         Wichita Falls
##                    60
```

```
dim(data)[1]
```

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

```
prob_beaumont= 60/240
prob_beaumont
```

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

*la probabilità che esca la città di Beaumont è del 25%.*

```
#probabilità Mese Luglio

table(month)
```

```
## month
##  1  2  3  4  5  6  7  8  9 10 11 12
## 20 20 20 20 20 20 20 20 20 20 20 20
```

```
dim(data)[1]
```

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

```
prob_july= 20/240
prob_july
```

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

*la probabilità che esca il mese di luglio è del 8,3%.*

```
#probabilità Mese dicembre 2012

dataframe= data.frame (anno=c(year),
                       mese=c(month))

subset_dec2012 <- subset(dataframe, anno == 2012 & mese == 12)
subset_dec2012
```

```
##     anno mese
## 36  2012   12
## 96  2012   12
## 156 2012   12
## 216 2012   12
```

```
dim(data)
```

```
## [1] 240   8
```

```
prob_dec2012=4/240
prob_dec2012
```

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

*il mese di dicembre anno 2012 si presenta 4 volte nel dataset e la probabilità che esca è del 1,6%.*

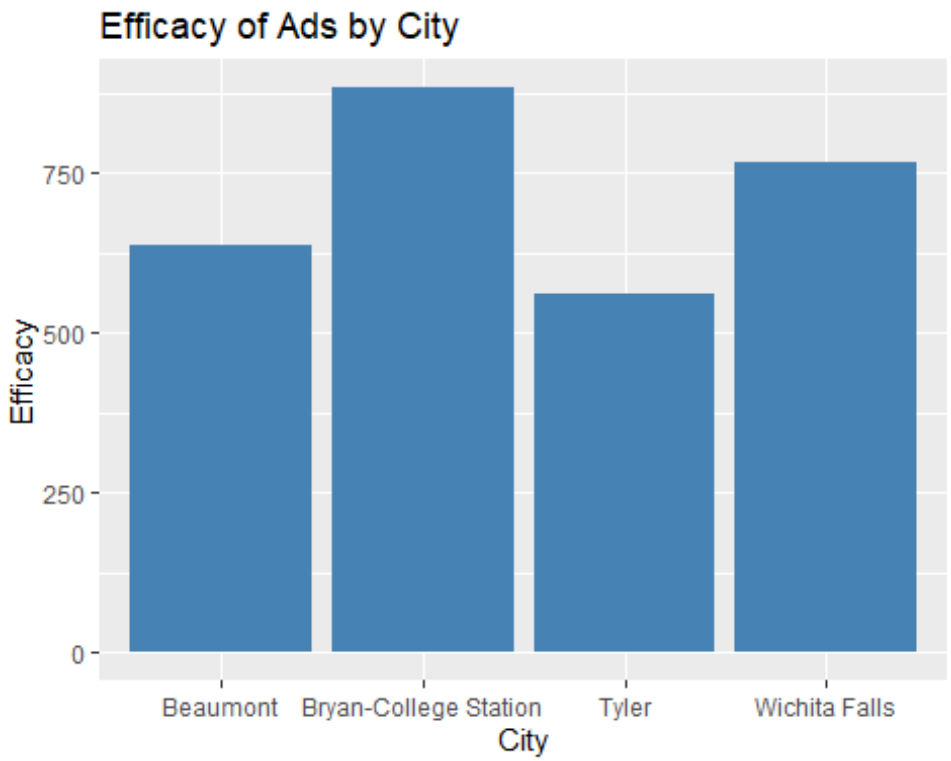
8) Esiste una colonna col prezzo mediano, creane una che indica invece il prezzo medio, utilizzando le altre variabili che hai a disposizione

```
data$mean_price=(data$volume/data$sales)
```

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

*metto in relazione sales e listings per indicare l'efficacia di vendita rispetto agli annunci attivi.*

```
data$efficacy <- data$sales / data$listings * 100

library(ggplot2)

Efficacy_plot=ggplot(data, aes(x = city, y = efficacy)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  xlab("City") +
  ylab("Efficacy") +
  ggtitle("Efficacy of Ads by City")

Efficacy_plot
```

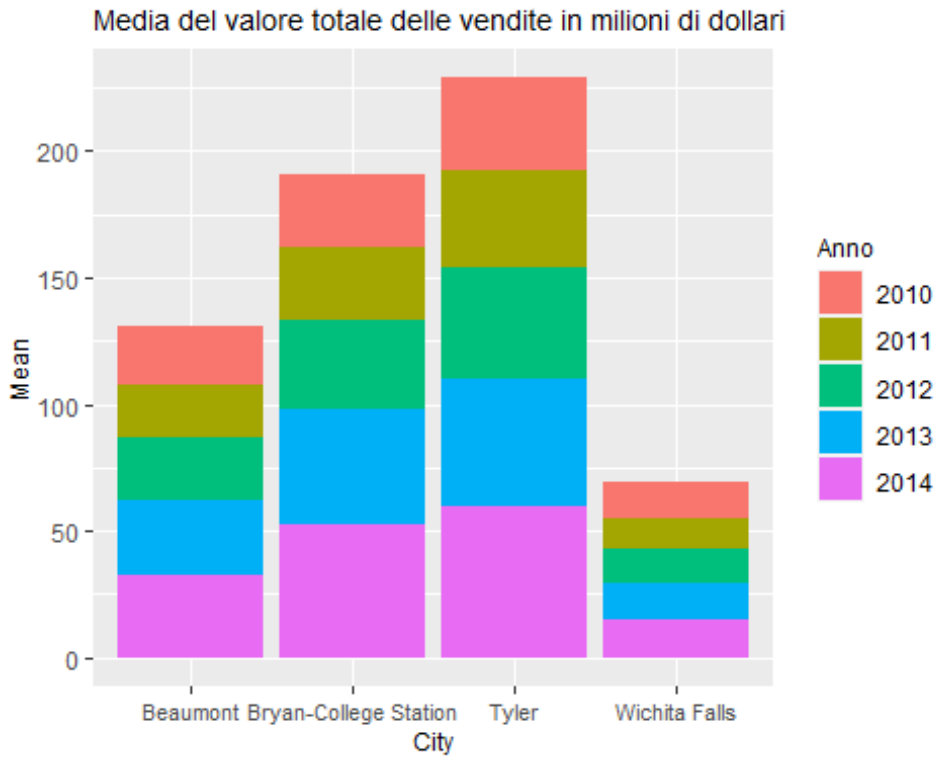



*Il grafico mostra l'efficacia degli annunci per le vendite in ciascuna città. Un valore di efficacy alto potrebbe indicare che gli annunci attivi sono efficaci nel generare vendite. Al contrario, un valore basso potrebbe indicare che, nonostante ci siano molti annunci attivi, le vendite restano basse e va migliorato qualcosa. Nel caso riportato la maggiore efficacia la si riscontra negli annunci per la città di Bryan-college Station.*

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.

```
install.packages("dplyr")

library(dplyr)

data %>%
  group_by(city, year) %>%
  summarise(media=mean(volume),
            dev.std=sd(volume))

grafico_dplyr=data %>%
  group_by(city, year) %>%
  summarise(Media = mean(volume), dev_standard = sd(volume)) %>%
  ggplot(aes(x = city, y = Media, fill = as.factor(year))) +
  geom_bar(stat = "identity") +
  ggtitle("Media del valore totale delle vendite in milioni di dollari")+
  labs(x="City", y="Mean", fill="Anno")+
  theme(axis.text.x = element_text(size=8))+
  theme(title = element_text(size=9))
```

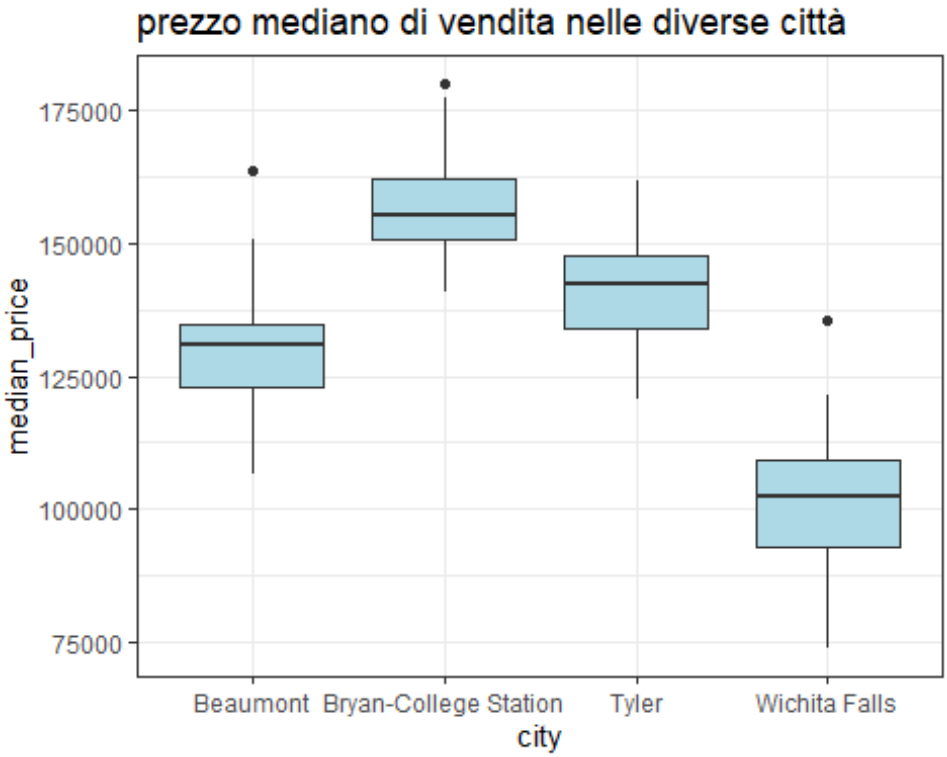
```
grafico_dplyr
```




*Il grafico indica la media del valore totale delle vendite in milioni di dollari per ciascuna città in base all'anno di interesse. L'asse delle x riporta la variabile city, l'asse delle y la media del valore totale delle vendite. Quest'ultima viene calcolata per ogni città in base agli anni di riferimento, riportati nella legenda con diversi colori. Il maggior guadagno sembra concentrarsi nella città di Tyler nell'anno 2014.*

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

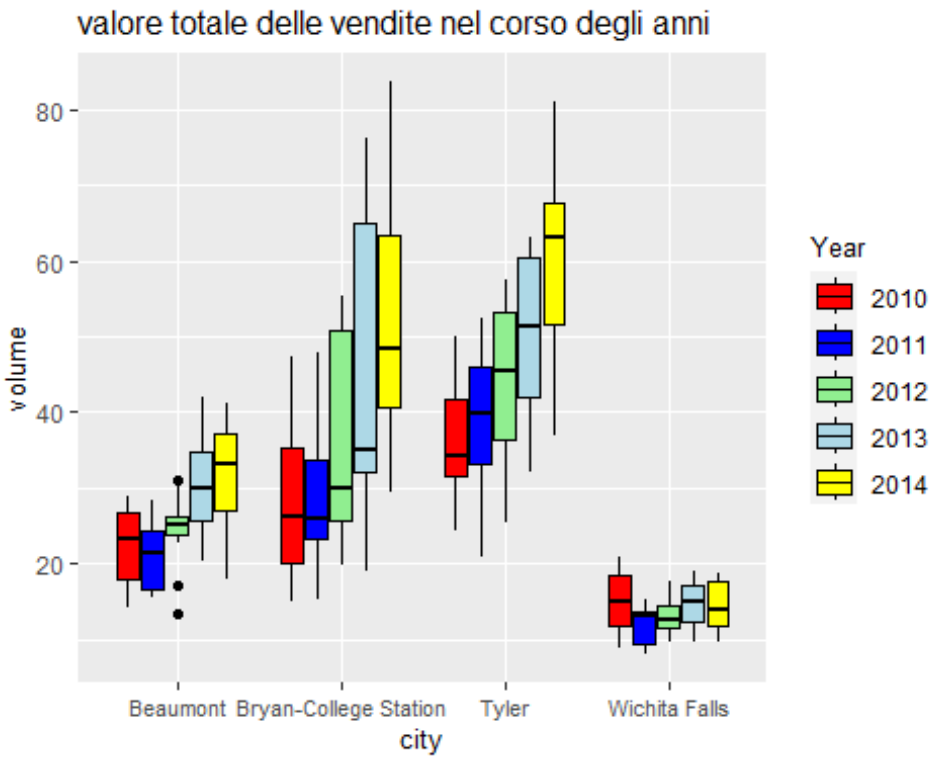
```
library(ggplot2)



mp <- ggplot(data = data) +
  geom_boxplot(aes(x = city,
                   y = median_price),
               fill = "lightblue") +
  theme_bw() +
  labs(title = "prezzo mediano di vendita nelle diverse città")

print(mp)
```




*Il grafico di bloxplot rappresenta il valore mediano di vendita riportato sull'asse y, nelle città indicate sull'asse x. I punti al di fuori del boxplot rappresentano gli outlier che sono presenti per tutte le città tranne che per la città di Tyler. Inoltre, nella città di Wichita Falls il prezzo mediano di vendita è più basso che nelle altre città.*

10.2) 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?

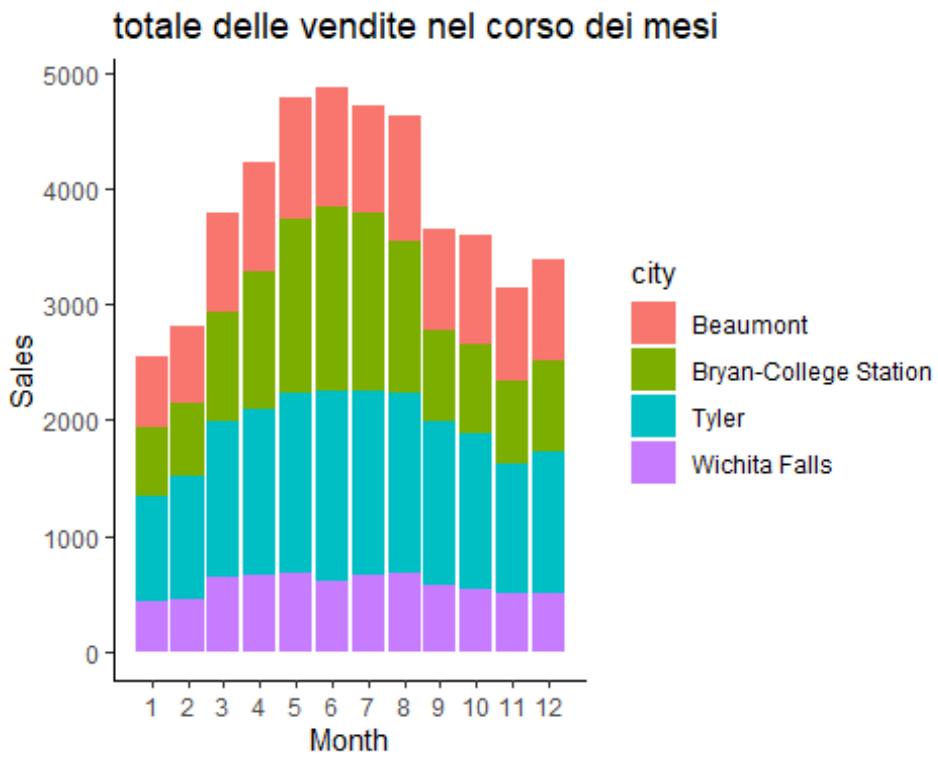
```
library(ggplot2)

Volume_boxplot=ggplot(data = data) +
  geom_boxplot(aes(x = city,
                   y = volume,
                   fill = as.factor(year)),
               color = "black") +
  scale_fill_manual(values = c("2010" = "red", "2011" = "blue", "2012" = "lightgre
en", "2013"= "lightblue","2014"="yellow")) +
  labs(fill = "Year", title = "valore totale delle vendite nel corso degli anni")+
  theme(axis.text.x = element_text(size=8), title=element_text(size=10))

print(Volume_boxplot)
```




*Il grafico mostra il valore totale delle vendite nel corso degli anni nelle 4 città. In particolare, per ogni città indicata sull'asse x, viene riportato il valore totale delle vendite per ciascun anno, rappresentato come boxplot di colore diverso (vedi legenda). Le maggiori vendite sembrano verificarsi nelle città di college-station e Tyler, sopratutto negli anni 2013-2014.*

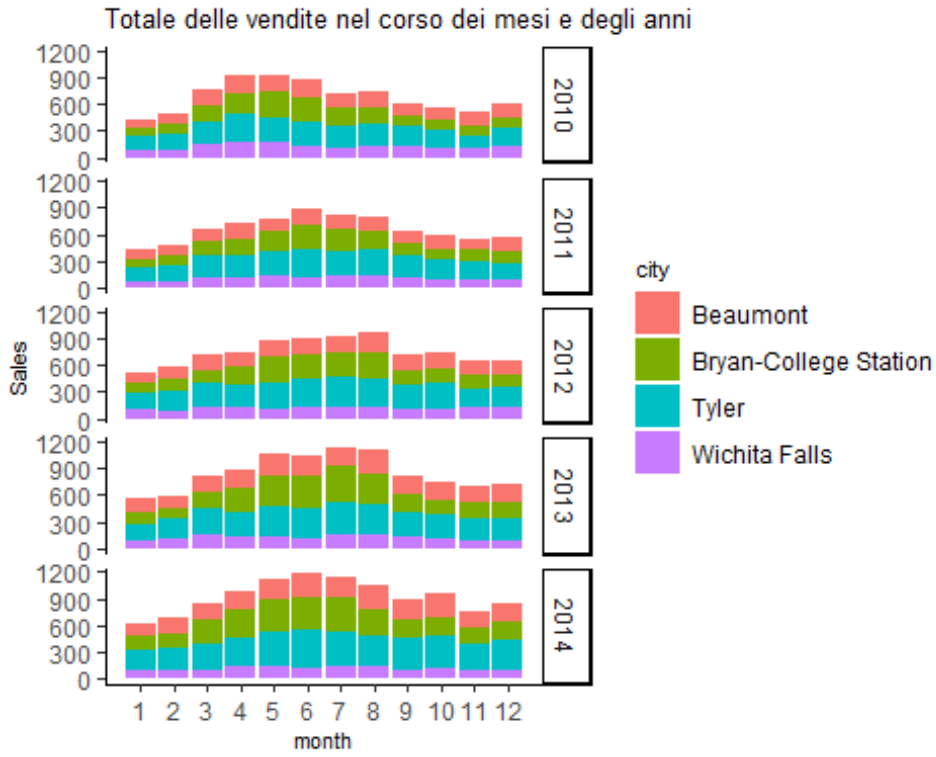
10.3) Usa un grafico a barre sovrapposte per confrontare il totale delle vendite nei vari mesi, sempre considerando le città. Prova a commentare ciò che viene fuori. Già che ci sei prova anche il grafico a barre normalizzato. Consiglio: Stai attento alla differenza tra geom_bar() e geom_col(). PRO LEVEL: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

```
#grafico senza variabile year

library(ggplot2)

sales_plot=ggplot(data=data)+
  geom_bar(aes(x = month, y = sales, fill=city),
           position="stack",
            stat = "identity") +

  xlab("Month") +
  ylab("Sales") +
  ggtitle("totale delle vendite nel corso dei mesi")+
  theme(title=element_text(size=8))+
  scale_x_continuous(breaks=seq(1,12,1))+
  theme_classic()

sales_plot
```



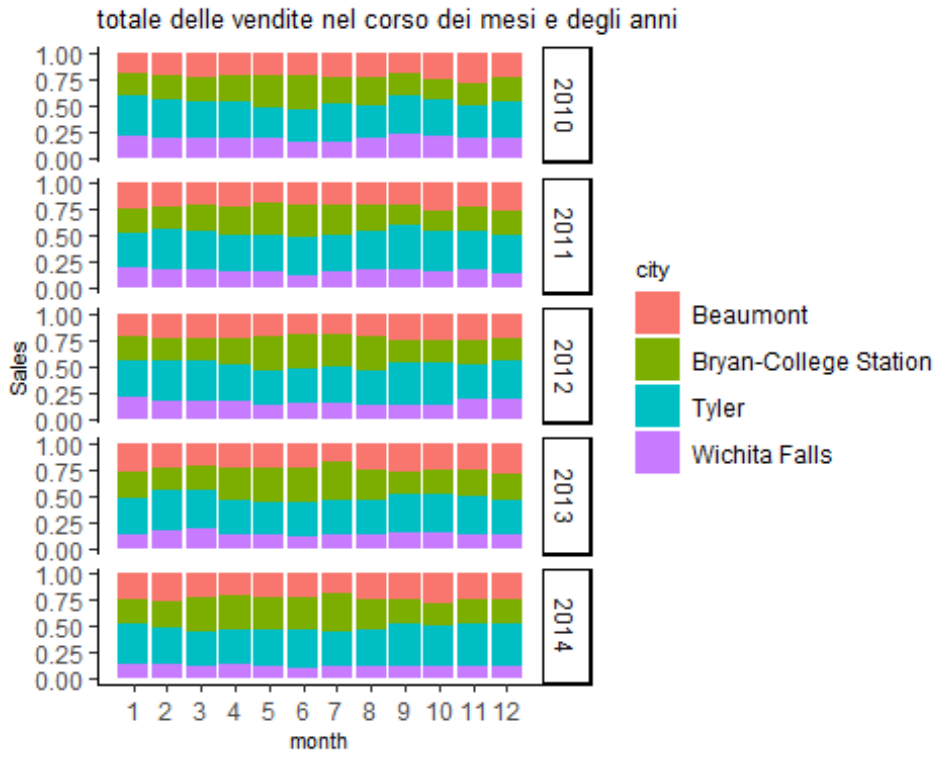

*Il grafico rappresenta il totale delle vendite per ciascuna città nel corso dei mesi. I mesi sono riportati sull'asse delle x, mentre il totale delle vendite sull'asse delle y. I diversi colori rappresentano le 4 città come riportato in legenda. Dal grafico così ottenuto possiamo concludere che vendite maggiori si accumulano nei mesi centrali, soprattutto per le città di Bryan College station e Tyler.*

```
#Aggiungo variabile YEAR utilizzando la funzione facet_grid() che mi divide il gra
fico in più pannelli

sales_plot=ggplot(data=data)+
  geom_bar(aes(x = month, y = sales, fill=city),
           position="stack",
            stat = "identity") +

  xlab("month") +
  ylab("Sales") +
  ggtitle("Totale delle vendite nel corso dei mesi e degli anni")+
  scale_x_continuous(breaks=seq(1,12,1))+
  theme_classic()+
  theme(title=element_text(size=8))+
  facet_grid(year)

sales_plot
```




*Rispetto al grafico precedente, qui è stata aggiunta la variabile Year in modo da valutare il totale delle vendite per ogni mese e ogni città dal 2010 al 2014. Possiamo ora concludere che le vendite maggiori si accumulano nei mesi centrali, nell'anno 2014 per la città di Tyler e Bryan-College Station. Anche per la città di Beaumont si osserva un leggero incremento delle vendite nei mesi centrali del 2013 e 2014. La città di Wichita Falls mantiene invece un basso numero di vendite in tutti gli anni di misurazione.*

```
#normalizzo il grafico

sales_plot=ggplot(data=data)+
  geom_bar(aes(x = month, y = sales, fill=city),
           position="fill",
            stat = "identity") +

  xlab("month") +
  ylab("Sales") +
  ggtitle("totale delle vendite nel corso dei mesi e degli anni")+
  scale_x_continuous(breaks=seq(1,12,1))+
  scale_y_continuous(breaks=seq(0,1,0.25))+
  theme_classic()+
   theme(title=element_text(size=8))+
  facet_grid(year)

sales_plot
```

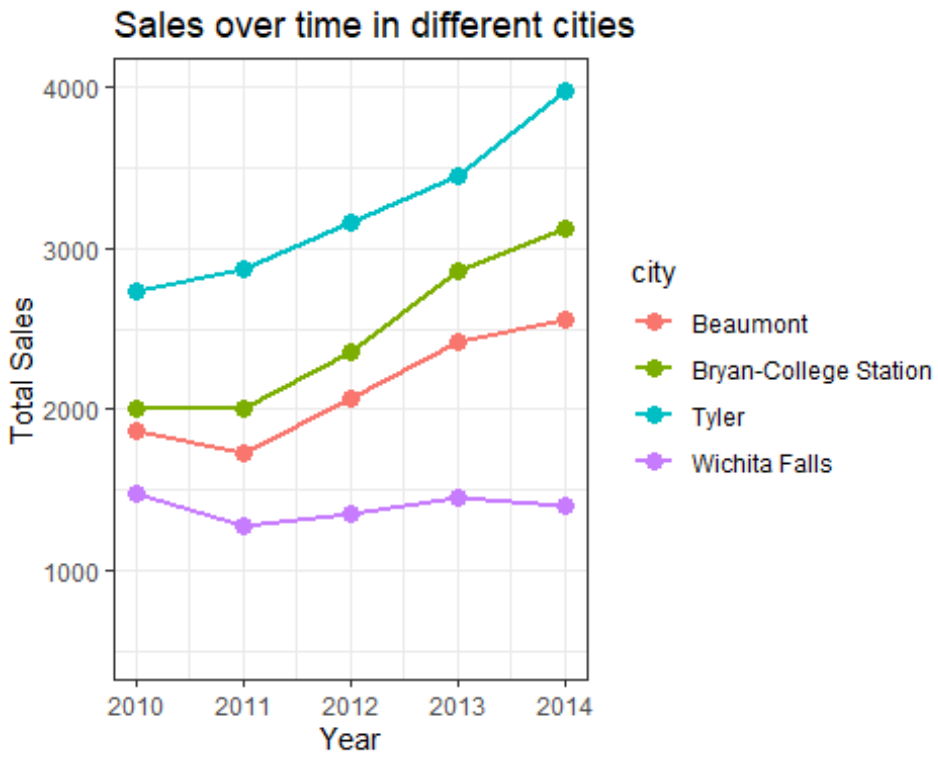



*Questo grafico normalizzato mette in relazione il totale delle vendite nel tempo di 4 città riportate in legenda. Rispetto ai grafici precedenti, possiamo confermare che si ha un incremento delle vendite nei mesi centrali per la città di Tyler e Bryan-College Station a prescindere dall'anno di osservazione. Al contrario, la città di Beaumont sembra ridurre le vendite nei mesi centrali e l'andamento delle vendite sembra seguire lo stesso trend per tutti gli anni. La città di wichita falls, oltre ad avere un numero basso di vendite rispetto alle altre città ha anche un calo delle vendite nel 2014 rispetto agli anni precedenti.*

10.4) Prova a creare un line chart di una variabile a tua scelta per fare confronti commentati fra città e periodi storici. Consigli: Prova inserendo una variabile per volta. Prova a usare variabili esterne al dataset, tipo vettori creati da te appositamente.

```
library(dplyr)
library(ggplot2)

#creo vettore città

cities=c("Beaumont", "Bryan-College Station", "Tyler","Wichita Falls")

#uso dplyr per avere i dati di sales sommati in ciascun anno e per ciascuna città

data_summary <- data %>%
  filter(city %in% cities ) %>%
  group_by(city, year) %>%
  summarise(totale = sum(sales))

## `summarise()` has grouped output by 'city'. You can override using the
## `.groups` argument.
```

```r
# grafico con ggplot2 utilizzando come dati quelli ottenuti da dplyr

grafico_dplyr <- ggplot(data = data_summary) +
  geom_line(aes(x = year, y = totale, col = city), lwd = 1) +
  geom_point(aes(x = year, y = totale, col = city), size = 3) +
  labs(title = "Sales over time in different cities",
       x = "Year",
       y = "Total Sales") +
  theme(title=element_text(size=8))+
  scale_y_continuous(limits = c(500,4000))+
  theme_bw()

# mostra il grafico
print(grafico_dplyr)
```




*Le line chart indicano il totale delle vendite delle 4 città nel corso degli anni. Sull'asse delle x sono riportati gli anni dal 2010 al 2014, mentre sull'asse delle y il totale delle vendite in ciascun anno per ogni città. In questo modo è possibile confrontare l'andamento delle vendite tra le 4 città nel corso dei 5 anni. Osservando il grafico possiamo concludere che la città di Tyler, rispetto alle altre città, ha un maggior numero di vendite già dal primo anno di misurazione (2010) che continua a crescere fino al 2014. Al contrario, la città con un minor numero di vendite è Wichita Falls, dove le total sales si mantengono costanti dal 2010 al 2014.*