Final Project Statistica Descrittiva

Simone Santonoceto

2023-12-11

1) Scarica il dataset realestate_textas.csv da qui e importalo con R, questo contiene dei dati riguardanti le vendite di immobili in Texas. Le variabili del dataset sono:

```
realestate = read.csv("realestate_texas.csv", header = TRUE, sep = ",")
```

2) Indica il tipo di variabili contenute nel dataset.

```
head(realestate,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
```

```
attach(realestate)
```

City è una variabile qualitativa perchè indica la città in cui è stata effettuata la vendita Year e month le posso considerare variabili qualitative, nonostante siano dei numeri, perchè rappresentano l'anno e il mese in cui è stata effettuata la vendita sales e listings sono variabili quantitative discrete, in quanto esprimono un conteggio, non posso fare mezza vendita o pubblicare mezzo annuncio volume, median_price, e months_inventory sono variabili quantitative continue, esprimono una quantità, nonstante alcune non siano continue per loro natura, posso considerarle tali perchè hanno una risoluzione abbastanza elevata

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. Uso summary per calcolare gli indici di posizione, variabilità e forma per le variabili quantitative.

```
N_city <- length(city)
city_freq_abs <- table(city)
city_freq_rel <- city_freq_abs/N_city
city_distr_freq <- cbind(city_freq_abs, city_freq_rel)
city_distr_freq
```

```
##                       city_freq_abs city_freq_rel
## Beaumont                         60          0.25
## Bryan-College Station            60          0.25
## Tyler                            60          0.25
## Wichita Falls                    60          0.25
```

Ogni città ha la stessa frequenza assoluta e relativa, quindi la distribuzione di frequenza per la città è uniforme.

```
N_year <- length(year)
year_freq_abs <- table(year)
year_freq_rel <- year_freq_abs/N_year
year_distr_freq <- cbind(year_freq_abs, year_freq_rel)
year_distr_freq
```

```
##      year_freq_abs year_freq_rel
## 2010            48           0.2
## 2011            48           0.2
## 2012            48           0.2
## 2013            48           0.2
## 2014            48           0.2
```

Ogni anno ha la stessa frequenza assoluta e relativa, quindi la distribuzione di frequenza per gli anni è uniforme.

```
N_month <- length(month)
month_freq_abs <- table(month)
month_freq_rel <- city_freq_abs/N_month
month_distr_freq <- cbind(month_freq_abs, month_freq_rel)
month_distr_freq
```

```
##    month_freq_abs month_freq_rel
## 1              20           0.25
## 2              20           0.25
## 3              20           0.25
## 4              20           0.25
## 5              20           0.25
## 6              20           0.25
## 7              20           0.25
## 8              20           0.25
## 9              20           0.25
## 10             20           0.25
## 11             20           0.25
## 12             20           0.25
```

Ogni mese ha la stessa frequenza assoluta e relativa, quindi la distribuzione di frequenza per i mesi è uniforme.

```
summary(sales)
```

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

```
summary(volume)
```

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

```r
summary(median_price)
```

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

```r
summary(listings)
```

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

```r
summary(months_inventory)
```

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

4) Qual è la variabile con variabilità più elevata? Come ci sei arrivato? E quale quella più asimmetrica? Proviamo a calcolare i range di variazione assoluti delle variabili

```r
range <- function(x) {
  return (round(max(x)-min(x), digits = 3))
}

cat(
paste("Vendite ->",range(sales)),
paste("Volume ->",range(volume)),
paste("Prezzo mediano ->",range(median_price)),
paste("Annunci ->",range(listings)),
paste("Mesi medi per vendita ->",range(months_inventory)),
sep = "\n")
```

```
## Vendite -> 344
## Volume -> 75.381
## Prezzo mediano -> 106200
## Annunci -> 2553
## Mesi medi per vendita -> 11.5
```

Da una prima analisi sembrerebbe che la variabile che varia maggiormente sia il median price, tuttavia queste variabili hanno scale molto diverse ed è quindi bene calolare il coefficiente di variazione di ogni singola variabile

```r
cvar <- function(x) {
  return (round(sd(x)/mean(x)*100, digits = 3))
}

cat(
paste("Vendite ->",cvar(sales)),
paste("Volume ->",cvar(volume)),
paste("Prezzo mediano ->",cvar(median_price)),
paste("Annunci ->",cvar(listings)),
paste("Mesi medi per vendita ->",cvar(months_inventory)),
sep = "\n")
```

```
## Vendite -> 41.422
## Volume -> 53.705
## Prezzo mediano -> 17.082
## Annunci -> 43.308
## Mesi medi per vendita -> 25.06
```

Da questa analisi è evidente che la variabile con coefficiente di variazione più elevato sia infatti Volume e non median price, questo perchè la scala di prezzo mediano era in milioni di dollari ed il range era dunque più elevato

Utilizzo il pacchetto moments per calolare l'indice di asimmetria di fisher

```
library(moments)
cat(
paste("Vendite  ->", round(skewness(sales),3)),
paste("Volume  ->", round(skewness(volume),3)),
paste("Prezzo mediano ->", round(skewness(median_price),3)),
paste("Annunci  ->", round(skewness(listings),3)),
paste("Mesi medi per vendita  ->", round(skewness(months_inventory),3)),
sep = "\n")
```

```
## Vendite  -> 0.718
## Volume  -> 0.885
## Prezzo mediano -> -0.365
## Annunci  -> 0.649
## Mesi medi per vendita  -> 0.041
```

La variabile più asimmetrica è volume, perchè ha il valore assoluto di skewness più alto. Questa variabile ha asimmetria positiva, di fatti è verificata la relazione Media > Mediana > Moda. Visualizzo i dati

```
summary(volume)
```

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

Media e mediana confermano la relazione precedente, la moda essendo Volume una variabile quantitaiva va stimata, suddivido in classi la variabile volume e valuto la moda

```
volume_cl <- cut(volume, seq(5, 100, 10))
table(volume_cl)
```

```
## volume_cl
##  (5,15] (15,25] (25,35] (35,45] (45,55] (55,65] (65,75] (75,85] (85,95]
##      41      63      55      32      26      11       8       4       0
```

I valori più frequenti si concentrano sulle classi sotto a 25, quindi si conferma media > mediana > moda

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.

Divido la variabile sales in classi di ampiezza 50

```r
sales_cl <- cut(sales, seq(50, 450, 50))
table(sales_cl)
```

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

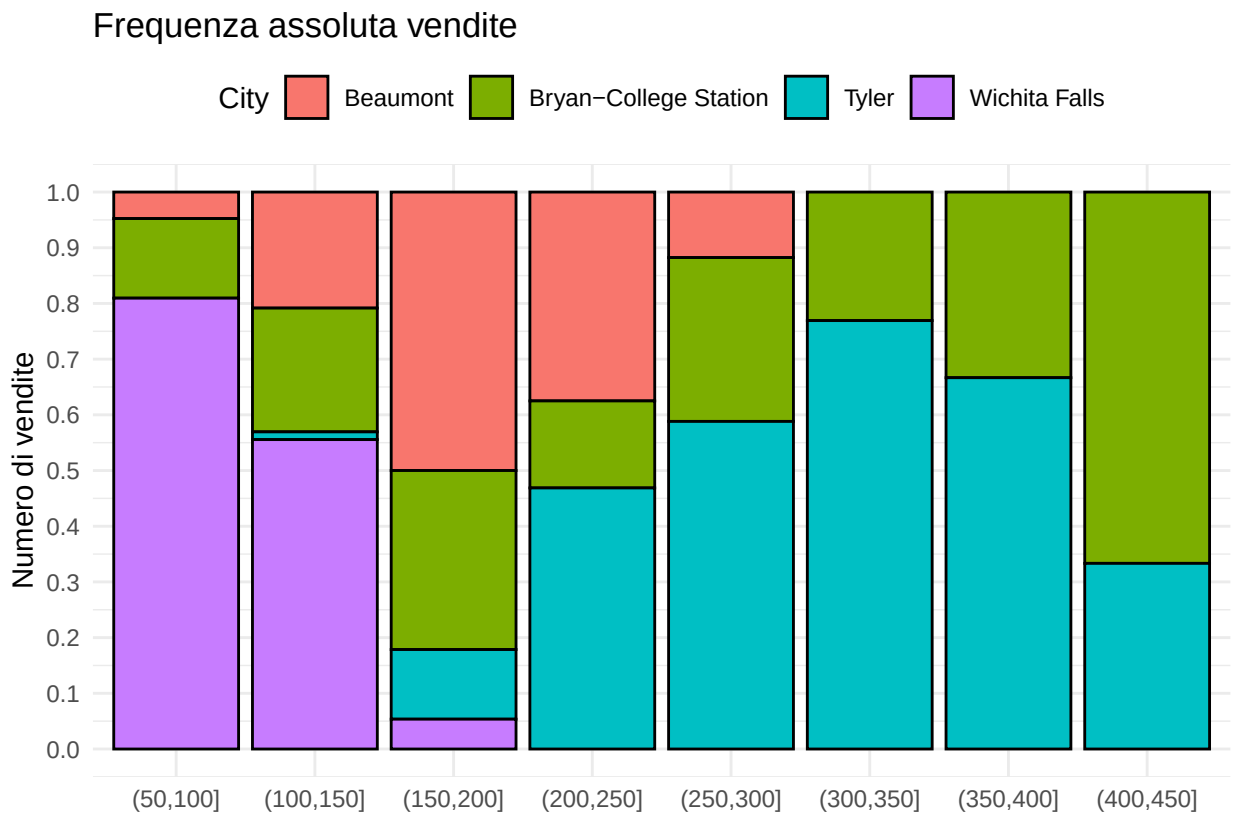
Costruisco la distribuzione di frequenza per la variabile sales

```r
N_sales_cl <- length(sales_cl)
sales_cl_freq_abs <- table(sales_cl)
sales_cl_freq_rel <- sales_cl_freq_abs/N_sales_cl
sales_cl_distr_freq <- cbind(sales_cl_freq_abs, sales_cl_freq_rel)
sales_cl_distr_freq
```

```
##           sales_cl_freq_abs sales_cl_freq_rel
## (50,100]                 21        0.08750000
## (100,150]                72        0.30000000
## (150,200]                56        0.23333333
## (200,250]                32        0.13333333
## (250,300]                34        0.14166667
## (300,350]                13        0.05416667
## (350,400]                 9        0.03750000
## (400,450]                 3        0.01250000
```

Costruisco il grafico a barre

```
library(ggplot2)
ggplot()+
  geom_bar(aes(x = sales_cl, fill = city), position = "fill",
           col = "black")+
  labs(x="", y="Numero di vendite", title = "Frequenza assoluta vendite")+
  scale_y_continuous(breaks = seq(0, 1, 0.1))+
  theme_minimal()+
  guides(fill=guide_legend(title="City"))+
  theme(legend.position = "top")
```



Dal grafico possiamo vedere che:

1) Withca Falls ha realizzato vendite solo nei range fino a 200k;
2) Beamount ha realizzato vendite nel range da 50k a 300k con vendite maggiori nel range 150k-250k;
3) Tyler è concentrata maggiormente nei range da 200k a 450k, con un picco nel range 300k-350k;
4) Bryan-College è distribuita in tutti i range e domina il range più alto 400k-450k.

Costruisco una funzione per calolare l'indice di Gini visto che dovrò usarla più volte

```
gini.index <- 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)
}
```

Calcolo l'indice di gini per la variabile sales suddivisa in classi

```
gini.index(sales_cl)
```

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

L'indice di Gini tende a 1, quindi la variabile sales è molto eterogenea

```
summary(sales_cl)
```

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

6) Indovina l'indice di gini per la variabile city e commenta il risultato.

La variabile city è equidistribuita, quindi l'indice di gini sarà 1 Calcoliamolo come conferma

```
gini.index(city)
```

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

Come supposto dall'analisi precedente, l'indice di gini è 1

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?

La distribuzione di frequenza è uniforme per queste 3 variabili: la probabilità che presa una riga a caso di questo dataset essa riporti la città "Beaumont" è 1/4 la probabilità che riporti il mese di Luglio è 1/12 la probabilità che riporti il mese di dicembre 2012 è: probabilità dicembre 1/12 * probabilità 2012 1/5 = 1/60

```
prob <- function(value) {
  N <- length(value)
  value_freq_abs <- table(value)
  value_freq_rel <- value_freq_abs/N
  value_distr_freq <- cbind(value_freq_rel)
  return(value_distr_freq)
}

prob(city)["Beaumont",1] == 1/4
```

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

```r
prob(month)["12",1] == 1/12
```

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

```r
prob(month)["12",1] * prob(year)["2012",1] == 1/5 * 1/12
```

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

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

Creo una colonna che indichi il prezzo medio usando il numero delle vendite ed il volume totale di vendita

```r
average_price <- volume/sales *1e6
realestate <- cbind(realestate, average_price)
head(realestate)
```

```
##       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
##   average_price
## 1      170626.5
## 2      163796.3
## 3      157697.8
## 4      134095.0
## 5      142737.6
## 6      144015.9
```

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

Avendo a disposizione il numero di annunci attivi ed il numero di vendite posso creare un indice di efficacia con il rapporto tra questi due valori

```r
efficacia <- sales/listings*100
realestate <- cbind(realestate, efficacia)
head(realestate)
```

```
##       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
##   average_price efficacia
```

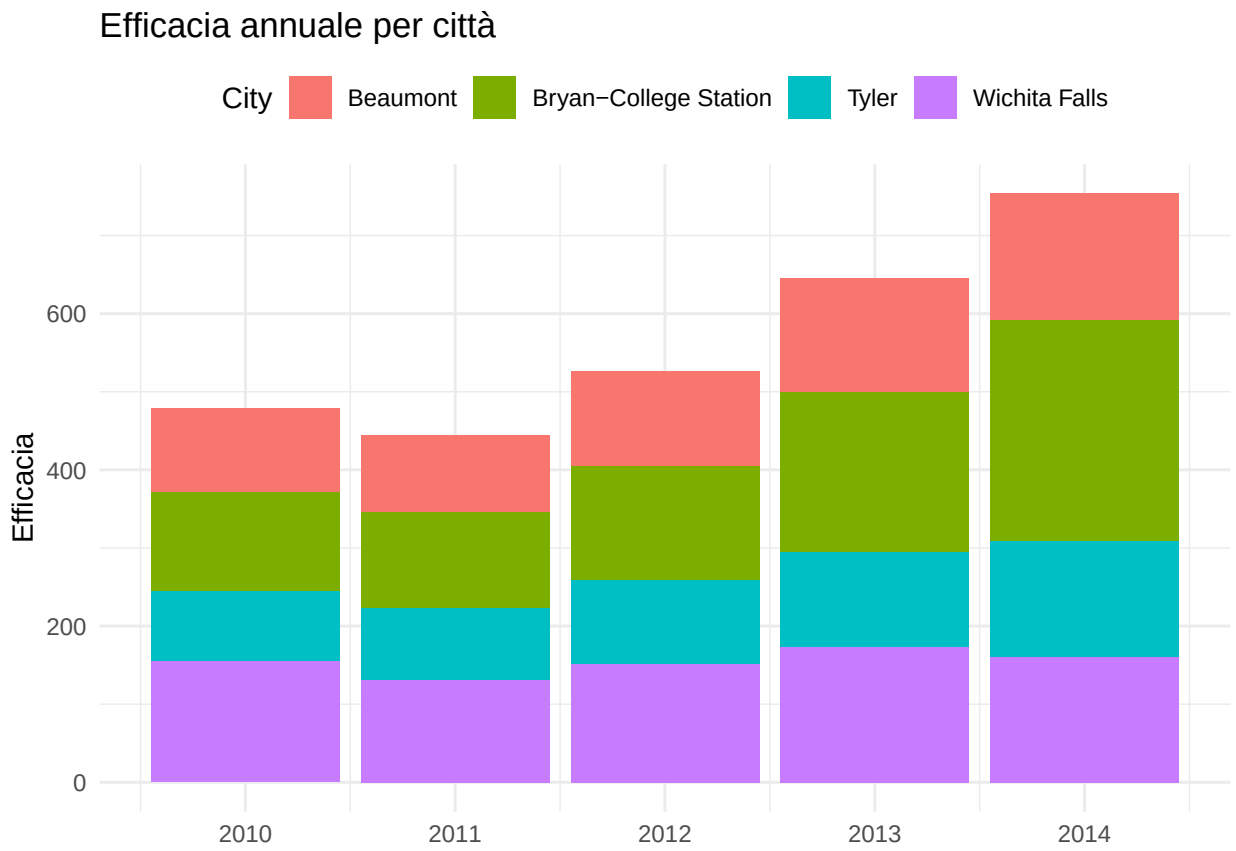
```
## 1      170626.5  5.414220
## 2      163796.3  6.809584
## 3      157697.8 10.775607
## 4      134095.0 11.709602
## 5      142737.6 11.405985
## 6      144015.9 10.482529
```

Per fare qualche considerazione raggruppo i dati per anno e vedo se ci sono stati anni migliori di altri

```
library(dplyr)
realestate %>%
  group_by(year) %>%
  summarise(efficacia_media = mean(efficacia))
```

```
## # A tibble: 5 x 2
##    year efficacia_media
##   <int>           <dbl>
## 1  2010            9.97
## 2  2011            9.27
## 3  2012           11.0
## 4  2013           13.5
## 5  2014           15.7
```

```
ggplot()+
  geom_bar(aes(x = year, y = efficacia, fill=city),stat = "identity")+
  labs(x=NULL, y="Efficacia", title = "Efficacia annuale per città")+
  theme_minimal()+
  guides(fill=guide_legend(title="City"))+
  theme(legend.position = "top")
```



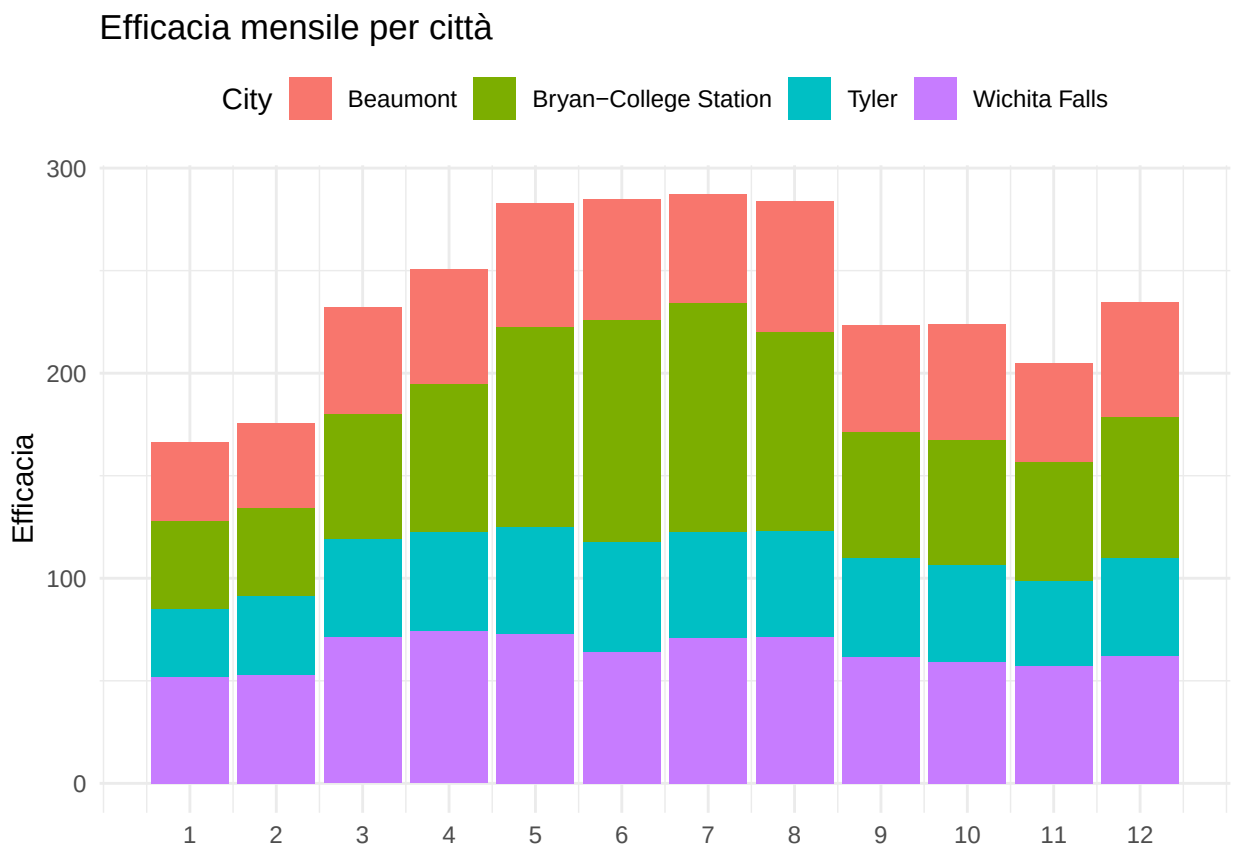
Da questi valori possiamo dedurre che il 2014 è stato l'anno dove gli annunci sono stati più efficaci, il 2011 è stato l'anno dove sono stati meno efficaci Nello specifico per Witchita Falls l'anno più efficace è stato il 2013, mentre il 2011 è stato l'anno meno efficacie per tutte le città

Effettuo lo stsso calcolo raggruppando per mese

```
realestate %>%
  group_by(month) %>%
  summarise(efficacia_media = mean(efficacia))
```

```
## # A tibble: 12 x 2
##    month efficacia_media
##    <int>           <dbl>
##  1     1            8.31
##  2     2            8.78
##  3     3           11.6
##  4     4           12.5
##  5     5           14.1
##  6     6           14.2
##  7     7           14.3
##  8     8           14.2
##  9     9           11.2
## 10    10           11.2
## 11    11           10.2
## 12    12           11.7
```

```r
ggplot()+
  geom_bar(aes(x = month, y = efficacia, fill = city), stat = "identity")+
  labs(x=NULL, y="Efficacia", title = "Efficacia mensile per città", position = "top")+
  scale_x_continuous(breaks = seq(1, 12, 1))+
  theme_minimal()+
  guides(fill=guide_legend(title="City"))+
  theme(legend.position = "top")
```



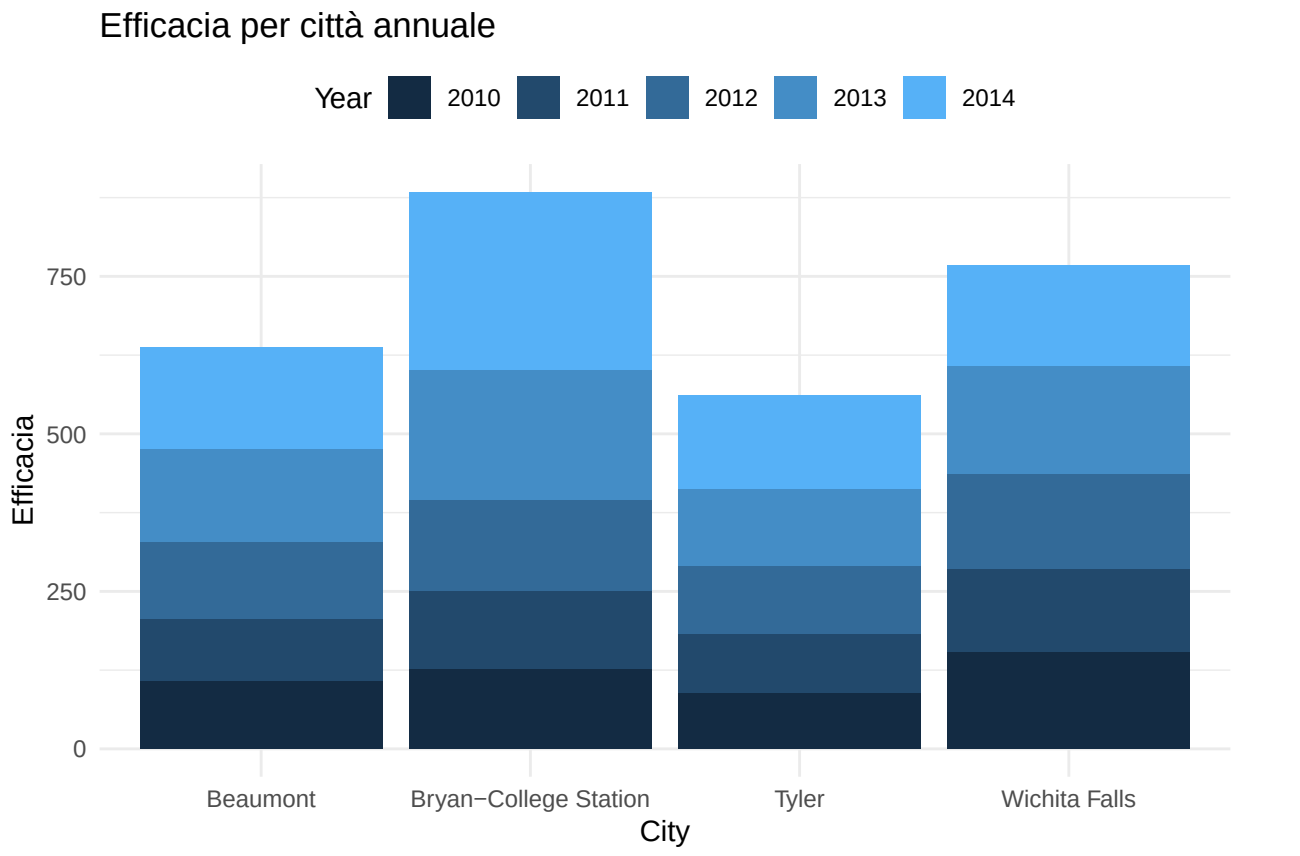
Da questi valori possiamo dedurre che i mesi dove gli annunci sono stati più efficaci sono quelli centrali e ciò sembra valido per tutte le città

Effettuo lo stsso calcolo raggruppando per città

```
realestate %>%
  group_by(city) %>%
  summarise(efficacia_media = mean(efficacia))
```

```
## # A tibble: 4 x 2
##   city                  efficacia_media
##   <chr>                           <dbl>
## 1 Beaumont                         10.6
## 2 Bryan-College Station            14.7
## 3 Tyler                            9.35
## 4 Wichita Falls                   12.8
```

```
ggplot()+
  geom_bar(aes(x = city, y = efficacia, fill = year), position = "stack", stat = "identity")+
  labs(x="City", y="Efficacia", title = "Efficacia per città annuale")+
  theme_minimal()+
  guides(fill=guide_legend(title="Year"))+
  theme(legend.position = "top")
```

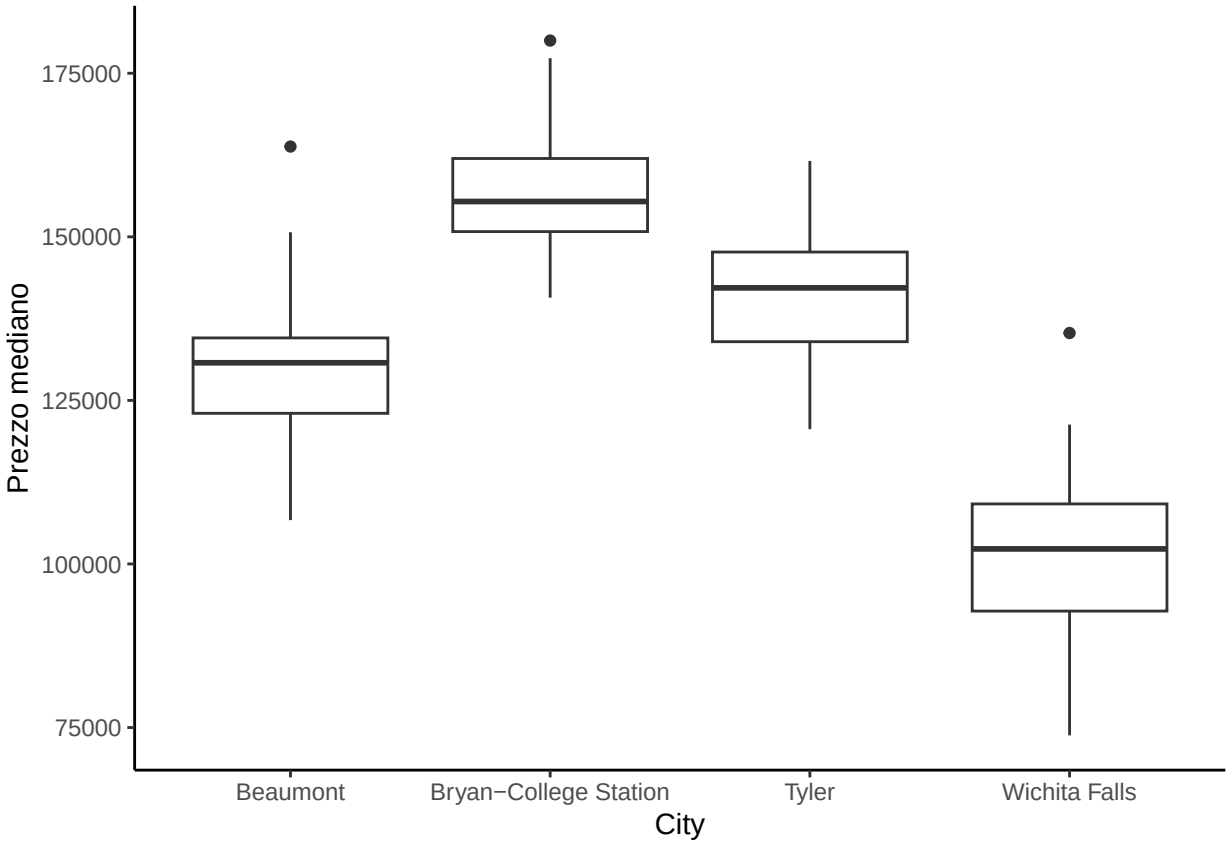


Da questi valori possiamo dedurre che la città con annunci più efficaci è Bryan-College Station, mentre la città con annunci meno efficaci è Tyler Inoltre si nota nuovamente l'efficacia per anno delle diverse città con le stesse conclusioni dell'analisi precedente

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. Riesci a fare qualche considerazione?

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

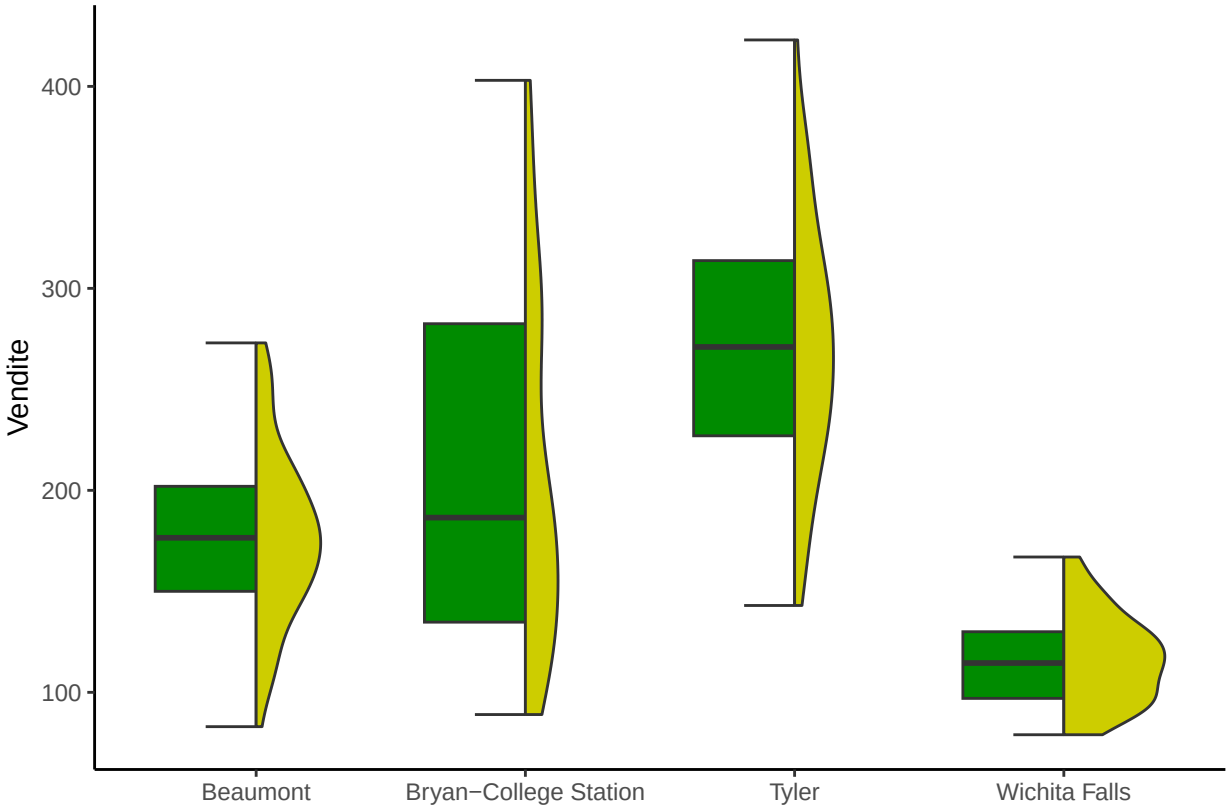
```
ggplot()+
  geom_boxplot(aes(y=median_price, x=city))+
  xlab("City")+
  scale_y_continuous("Prezzo mediano", seq(0,200e3,25e3))+
  theme_classic()
```



Bryan-College Station ha il prezzo mediano più alto, inoltre il suo primo quartile è più alto del terzo quartile di tutte le altre città Witcha Falls ha invece il prezzo mediano più basso di tutte e il terzo primo quartile è più basso del primo quartile di tutte le altre città Tyler, a differenza delle altre città non presenta nessun outlier Beaumont è quella che ha il prezzo mediano più centrato nell'intero range di prezzi mediani

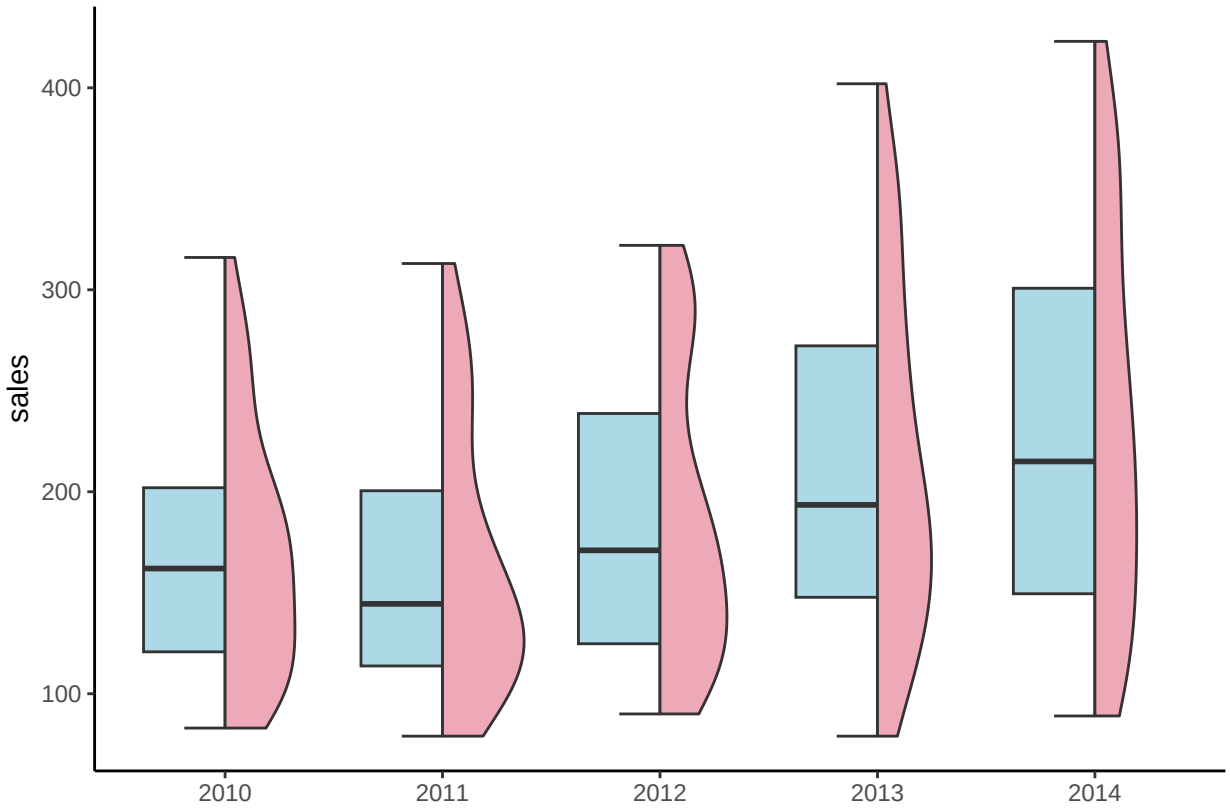
b) Utilizza i boxplot o qualche variante per confrontare la distribuzione del valore totale delle vendite tra le varie città ma anche tra i vari anni. Qualche considerazione da fare?

```
library(gghalves)

ggplot()+
  geom_half_boxplot(aes(x=city, y=sales), side="l", fill="green4")+
  geom_half_violin(aes(x=city, y=sales), side="r", fill="yellow3")+
  xlab("")+
  ylab("Vendite")+
  theme_classic()
```



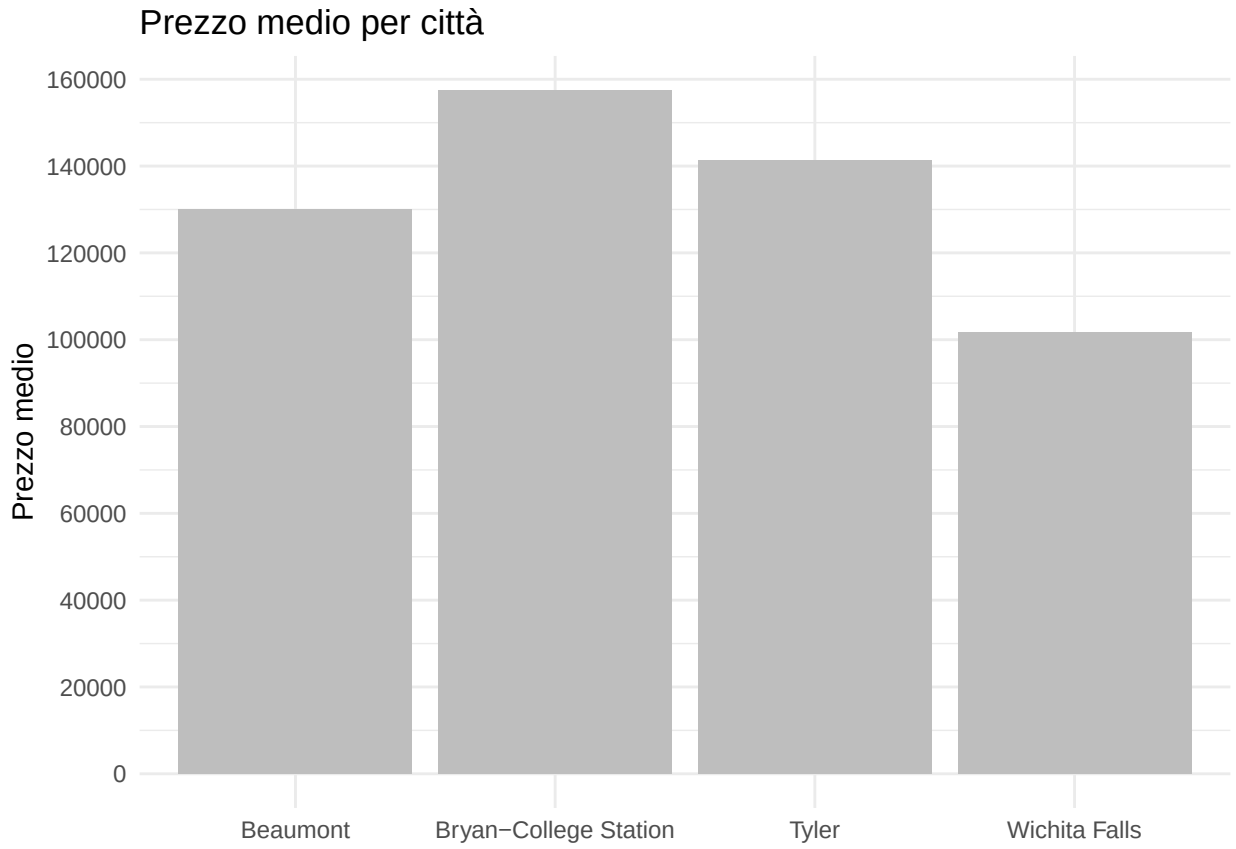
Da questo grafico possiamo notare come sono distribuite le vendite per città, in particolare: Bryan-College Station è asimmetrica negativa, Beaumont e tyler non presentano quasi asimmetria e Witcha falls è anch'essa asimmetrica negativa. Beamount e Witch Falls sono inoltre leptocurtiche come distribuzioni

```
library(gghalves)

ggplot()+
  geom_half_boxplot(aes(x=as.factor(year), y=sales), side="l", fill="lightblue")+
  geom_half_violin(aes(x=as.factor(year), y=sales), side="r", fill="pink2")+
  xlab("")+
  theme_classic()
```



Qui possiamo notare come il numero medio di vendite sti aumentando dai boxplot. Dal grafico a violino invece notiamo come le distibuzioni siano tutte asimmetriche negative e come le distribuzioni stiano passando da leptocurtiche a platicurtiche

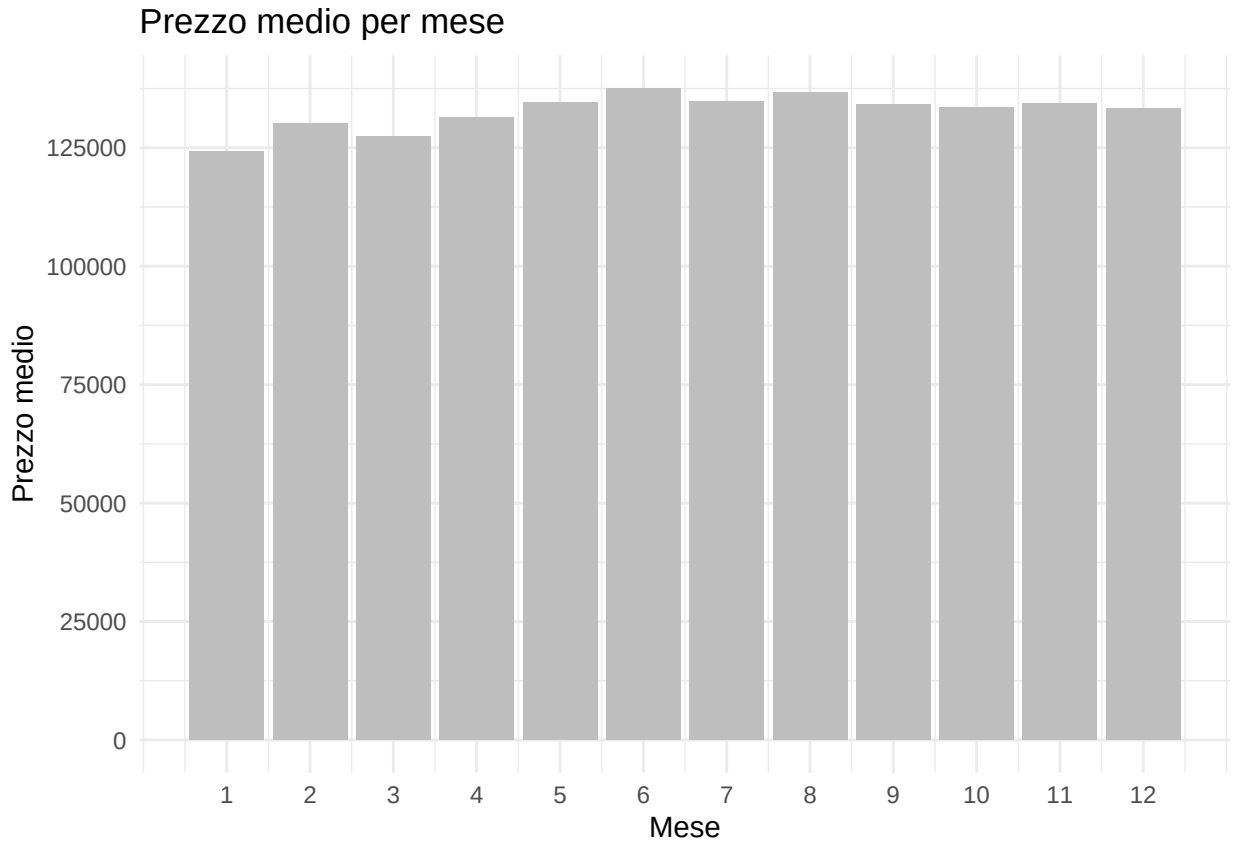
```r
realestate %>%
  group_by(city) %>%
  summarise(mean_median_price = mean(median_price), sd_price = sd(median_price),
            mean_sales = mean(sales), sd_sales = sd(sales),
            mean_listings = mean(listings), sd_listings = sd(listings),
            mean_volume = mean(volume), sd_volume = sd(volume),
            mean_average_price = mean(average_price), sd_average_price = sd(average_price),
            mean_efficacia = mean(efficacia), sd_efficacia = sd(efficacia)) %>%

ggplot()+
  geom_bar(aes(x = city, y = mean_median_price), fill="grey", position = "stack", stat = "identity")+
  labs(x=NULL, y="Prezzo medio", title = "Prezzo medio per città")+
  scale_y_continuous("Prezzo medio", seq(0,200e3,20e3))+
  theme_minimal()+
  guides(fill = "none")
```



Dal raggruppamento per città si vede la distribuzione del prezzo medio per ognuna di esse I prezzi medi sono nel range 100k-160k, Brian-College Station è al limte superiore e Witcha falls è verso il limite inferiore

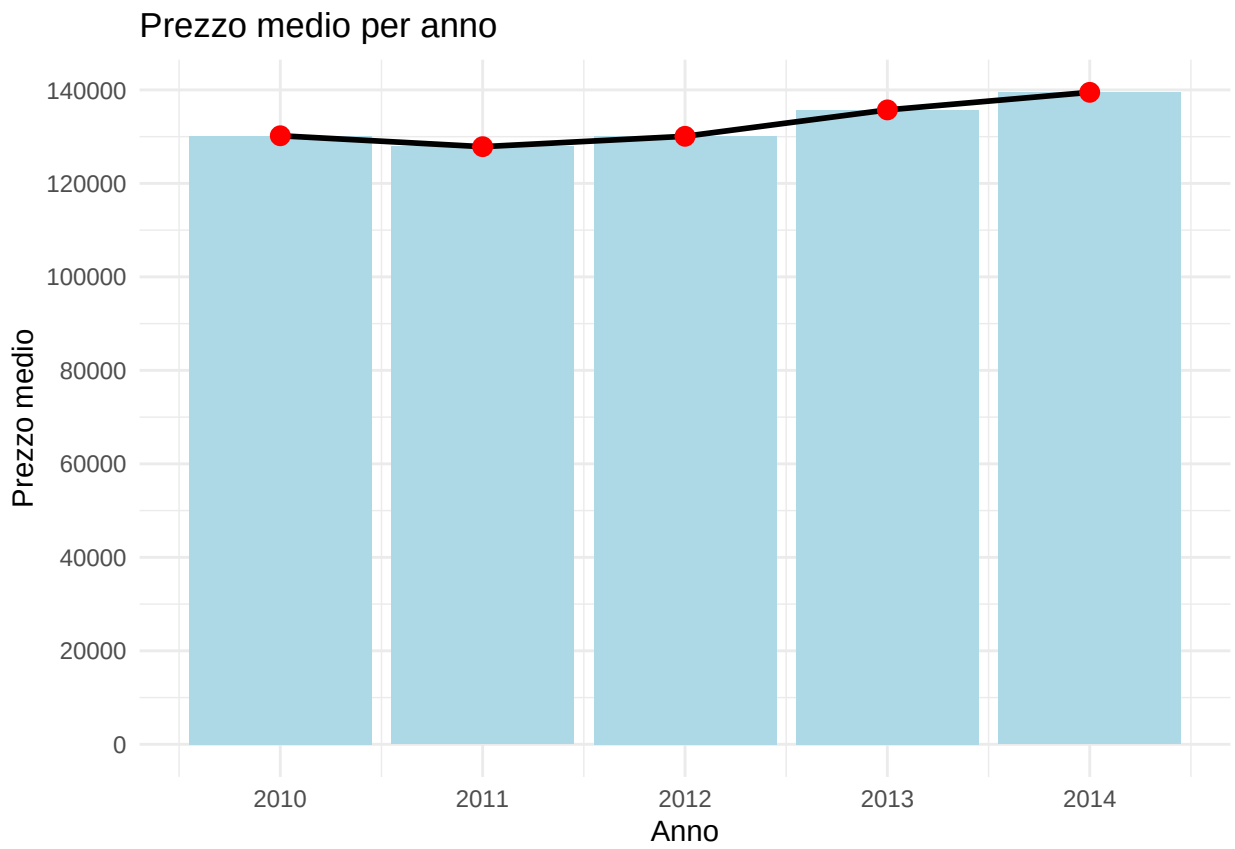
```r
realestate %>%
  group_by(month) %>%
  summarise(mean_median_price = mean(median_price), sd_price = sd(median_price),
            mean_sales = mean(sales), sd_sales = sd(sales),
            mean_listings = mean(listings), sd_listings = sd(listings),
            mean_volume = mean(volume), sd_volume = sd(volume),
            mean_average_price = mean(average_price), sd_average_price = sd(average_price),
```

```
            mean_efficacia = mean(efficacia), sd_efficacia = sd(efficacia)) %>%

ggplot()+
  geom_bar(aes(x = month, y = mean_median_price), fill="gray", position = "stack", stat = "identity")+
  labs(x=NULL, title = "Prezzo medio per mese")+
  scale_x_continuous("Mese", breaks = c(1:12))+
  scale_y_continuous("Prezzo medio", seq(0,200e3,25e3))+
  theme_minimal()+
  guides(fill = FALSE)
```



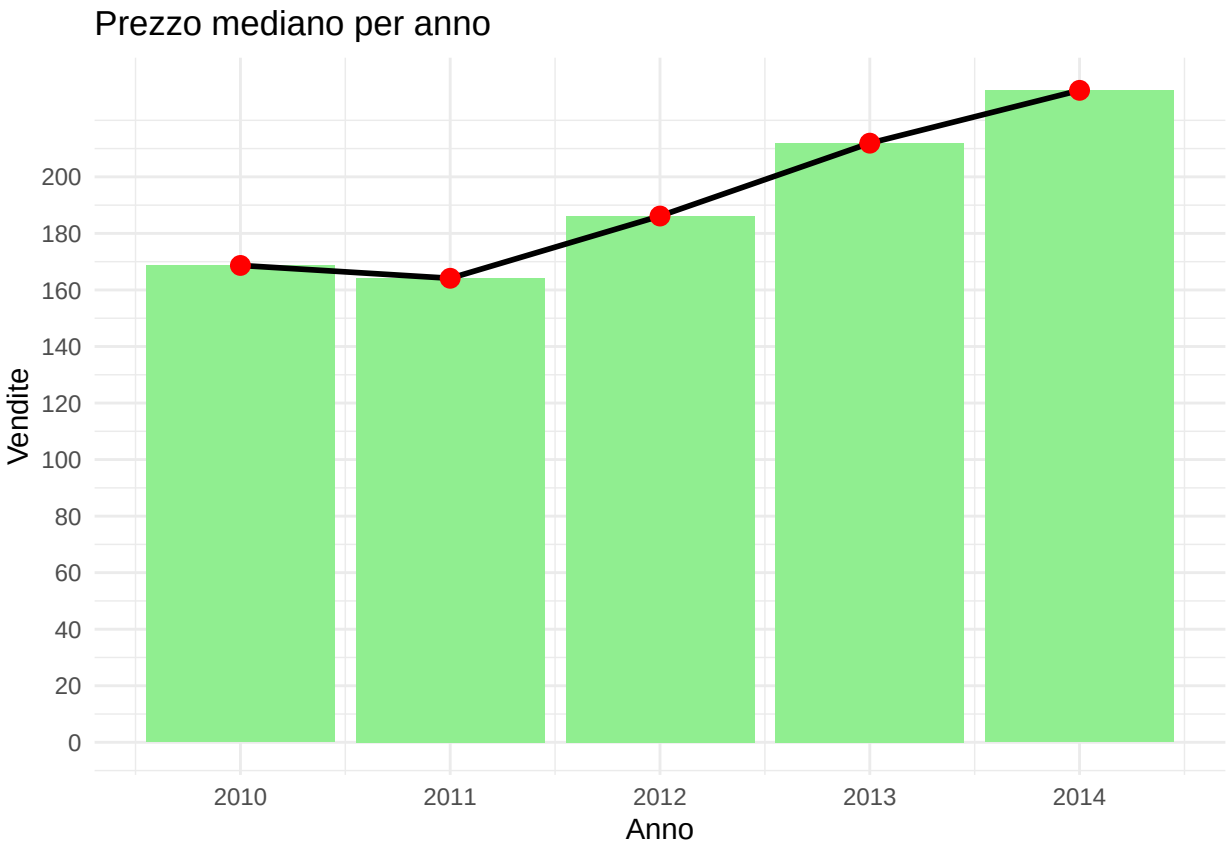
Il prezzo medio è abbastanza uniforme durante tutti i mesi

```
realestate %>%
  group_by(year) %>%
  summarise(mean_median_price = mean(median_price), sd_price = sd(median_price)) %>%

ggplot()+
  geom_bar(aes(x = year, y = mean_median_price), fill="lightblue", position = "stack", stat = "identity
  geom_freqpoly(aes(x = year, y = mean_median_price, group = 1), stat = "identity", color = "black", lw
  geom_point(aes(x = year, y = mean_median_price), size = 3, color = "red")+
  labs(x=NULL, y="Prezzo medio", title = "Prezzo medio per anno")+
  scale_x_continuous("Anno", breaks = c(2010:2014))+
  scale_y_continuous("Prezzo medio", breaks = seq(0, 200000, 20000))+
  theme_minimal()+
  guides(fill = FALSE)
```



Da questo grafico si apprezza il trend crescente del prezzo medio durante gli anni

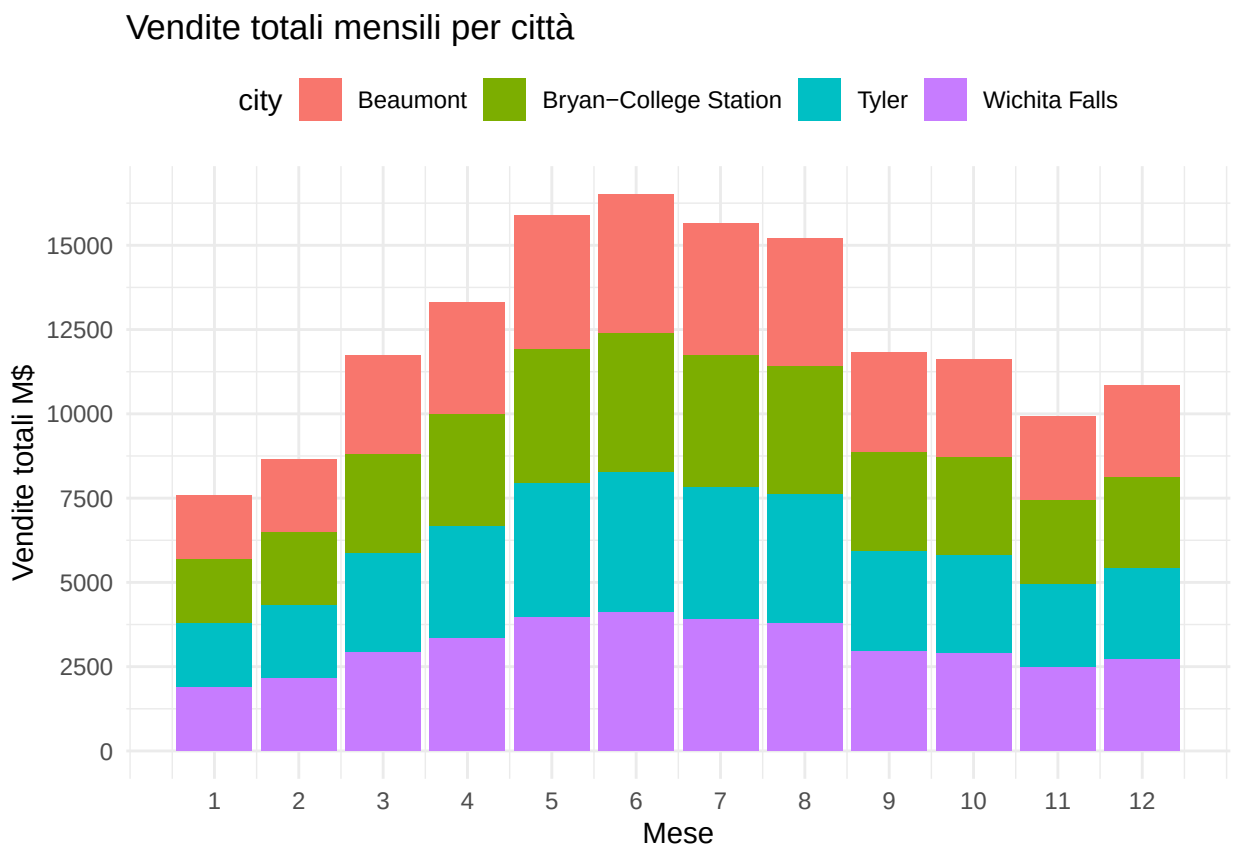
```
realestate %>%
  group_by(year) %>%
  summarise(mean_sales = mean(sales), sd_sales = sd(sales))%>%

ggplot()+
  geom_bar(aes(x = year, y = mean_sales), fill = "lightgreen", position = "stack", stat = "identity")+
  geom_freqpoly(aes(x = year, y = mean_sales, group = 1), stat = "identity", color = "black", lwd = 1)+
  geom_point(aes(x = year, y = mean_sales), size = 3, color = "red")+
  labs(x=NULL, y="Vendite", title = "Prezzo mediano per anno")+
  scale_x_continuous("Anno", breaks = c(2010:2014))+
  scale_y_continuous("Vendite", breaks = seq(0, 200, 20))+
  theme_minimal()+
  guides(fill = FALSE)
```



Si osserva che il prezzo mediano sta crescendo in maniera più evidente del prezzo medio, quindi anche se in media la distribuzione di prezzo cresce lentamente, il prezzo sta comunque crescendo ed il prezzo mediano ci fa apprezzare meglio questo fenomeno

c) Usa un grafico a barre sovrapposte per ogni anno, 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.

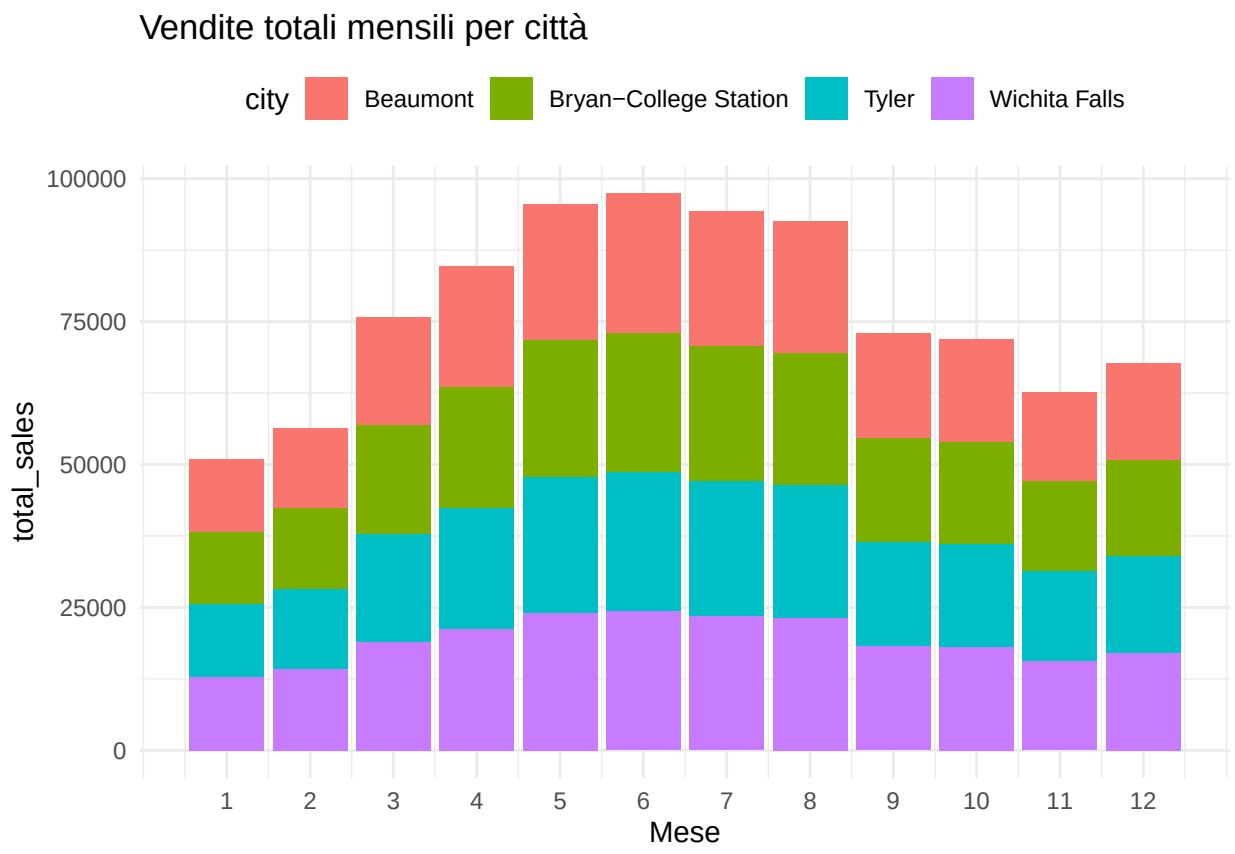
```
realestate %>%
  group_by(month) %>%
  reframe(total_sales = sum(sales*average_price), city=city)%>%

ggplot()+
  geom_bar(aes(x = month, y = total_sales/1e6, fill = city), stat = "identity")+
  labs(x=month, title = "Vendite totali mensili per città")+
  scale_x_continuous("Mese", breaks = c(1:12))+
  scale_y_continuous("Vendite totali M$", seq(0,20e3,25e2))+
  theme_minimal()+
  theme(legend.position = "top")
```



Risulta evidente come nei mesi centrali dell'anno le vendite aumentino, verifichiamo se perchè aumenta il numero o perchè aumentano i prezzi

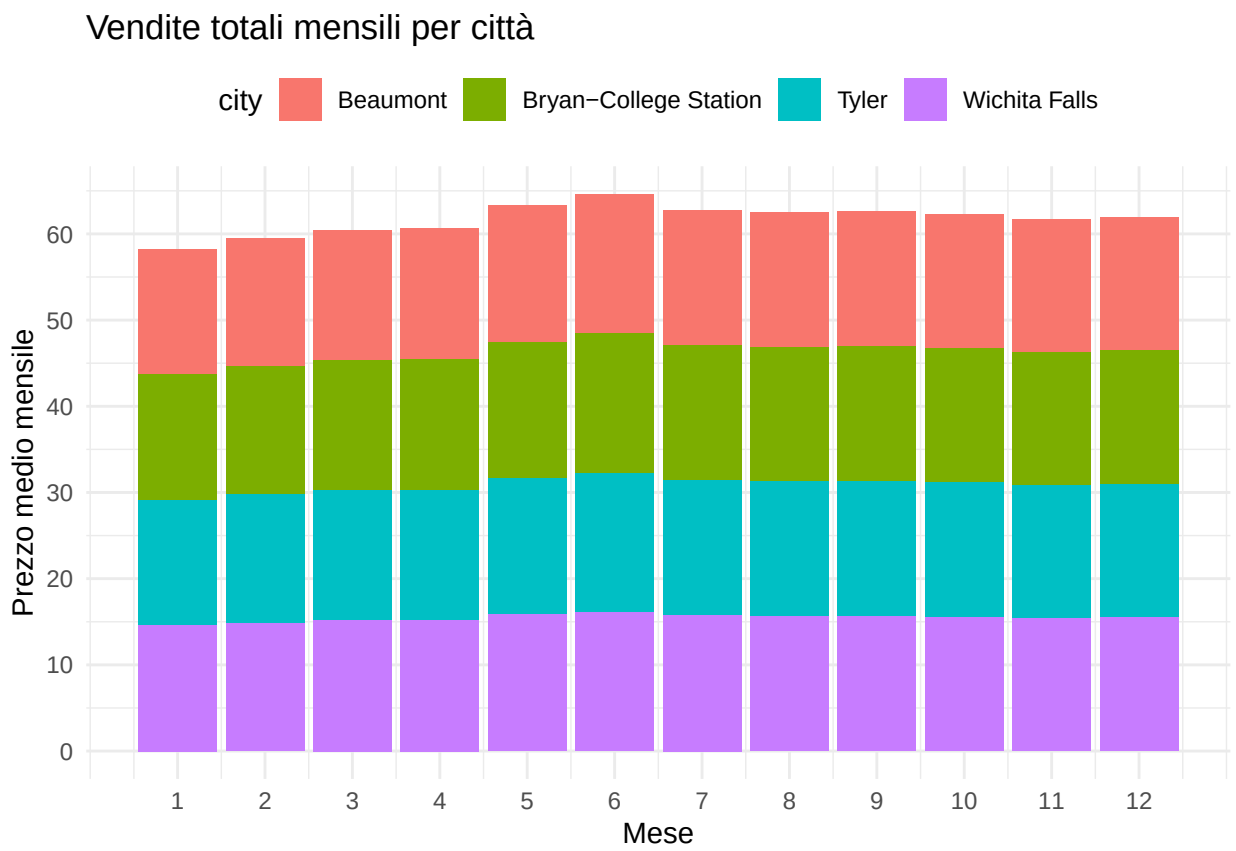
```
realestate %>%
  group_by(month) %>%
  reframe(total_sales = sum(sales), city=city)%>%

ggplot()+
```

```
  geom_bar(aes(x = month, y = total_sales, fill = city), stat = "identity")+
  labs(x=month, title = "Vendite totali mensili per città")+
  scale_x_continuous("Mese", breaks = c(1:12))+
  theme_minimal()+
  theme(legend.position = "top")
```



Il trend crescente nella parte centrale dell'anno è molto evidente

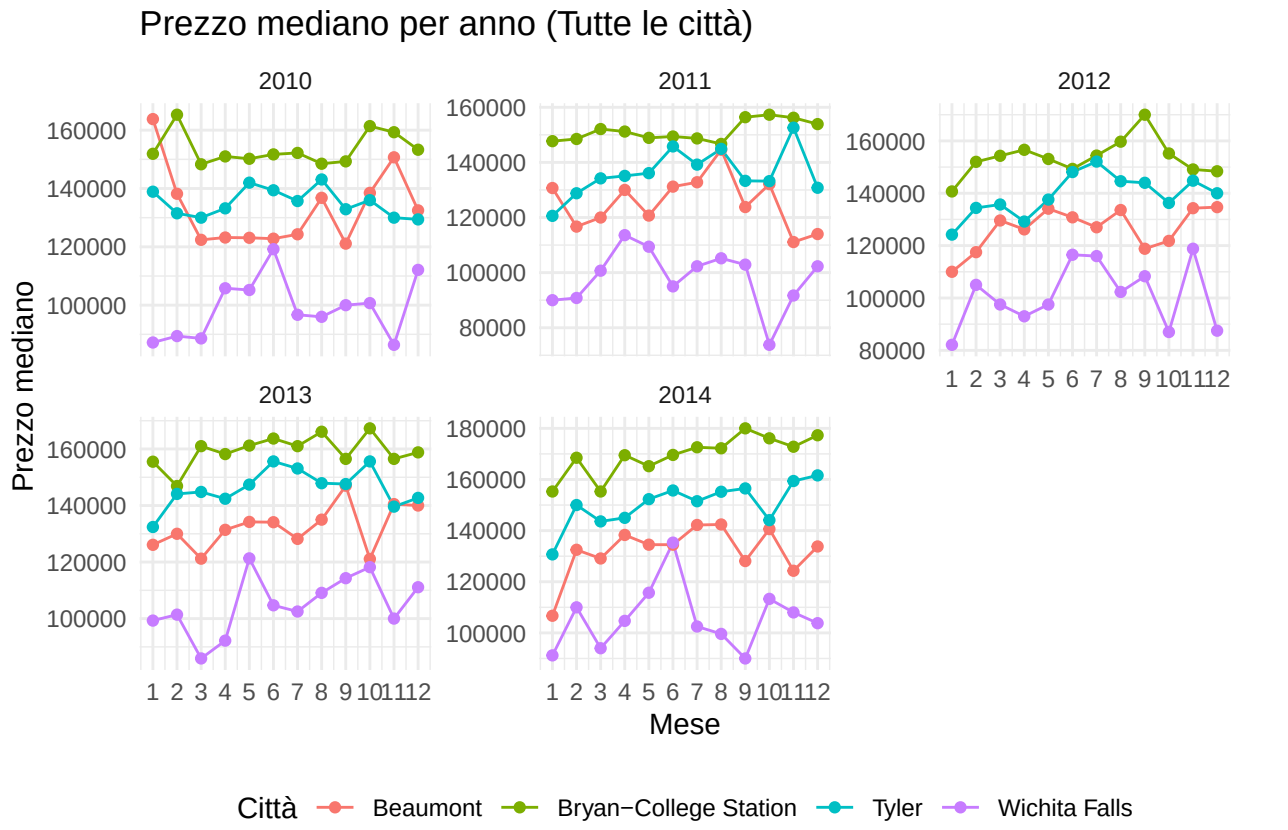
```
realestate %>%
  group_by(month) %>%
  reframe(sum_avg_price = sum(average_price), city=city)%>%

ggplot()+
  geom_bar(aes(x = month, y = sum_avg_price/1e6, fill = city), stat = "identity")+
  labs(x=month, title = "Vendite totali mensili per città")+
  scale_x_continuous("Mese", breaks = c(1:12))+
  scale_y_continuous("Prezzo medio mensile", seq(0,70,10))+
  theme_minimal()+
  theme(legend.position = "top")
```



Il trend è abbastanza costante durante tutti i mesi dell'anno, c'è un leggero incremento nella parte centrale Dunque il trend crescente nella parte centrale dell'anno è dato dall'aumento del prezzo medio, l'aumento del numero mensile di vendite amplifica il fenomeno

d) Crea 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.

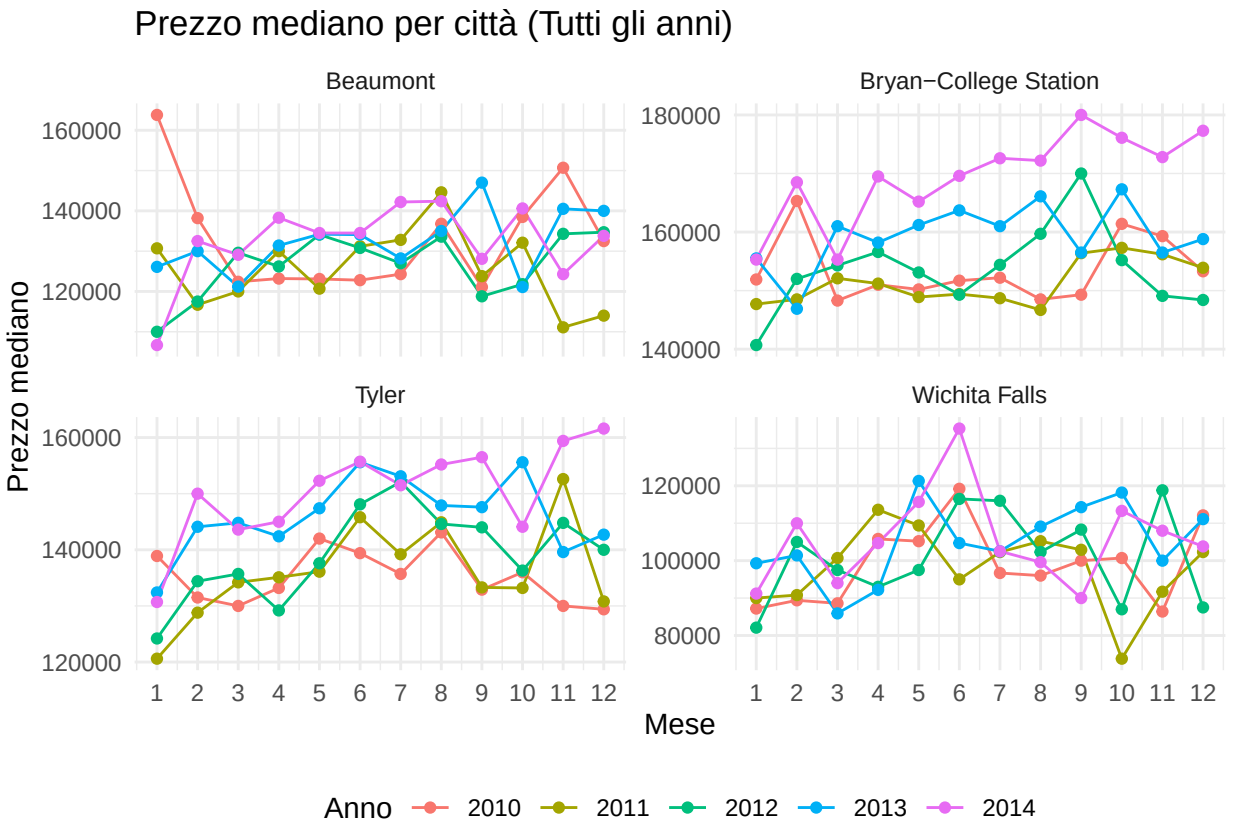
```
cities <- c("Beaumont", "Bryan-College Station", "Tyler", "Wichita Falls")
realestate_filtered <- filter(realestate, city %in% cities)

ggplot(realestate_filtered, aes(x = month, y = median_price, color =city)) +
  geom_line() +
  geom_point() +
  facet_wrap(~ year, scales = "free_y") +
  labs(x = NULL, y = "Prezzo mediano", title = "Prezzo mediano per anno (Tutte le città)") +
  scale_x_continuous("Mese", breaks = c(1:12)) +
  scale_y_continuous("Prezzo mediano", breaks = seq(0, 200000, 20000)) +
  scale_color_discrete(name = "Città") +
  theme_minimal() +
  theme(legend.position = "bottom")
```



Qui la visione d'insieme per ogni anno suddivisa per città e per mese. Si nota come i prezzi medi siano sempre più alti per Bryan-College Station.

```
cities <- c("Beaumont", "Bryan-College Station", "Tyler", "Wichita Falls")
realestate_filtered <- filter(realestate, city %in% cities)

ggplot(realestate_filtered, aes(x = month, y = median_price, color = as.factor(year))) +
```

```
  geom_line() +
  geom_point() +
  facet_wrap(~ city, scales = "free_y") +
  labs(x = "", y = "Prezzo mediano", title = "Prezzo mediano per città (Tutti gli anni)") +
  scale_x_continuous("Mese", breaks = c(1:12)) +
  scale_y_continuous("Prezzo mediano", breaks = seq(0, 200000, 20000)) +
  scale_color_discrete(name = "Anno") +
  theme_minimal() +
  theme(legend.position = "bottom")
```



Facendo un confronto per ogni città per ogni anno si può evincere il trend crescente per Bryan-College Station e Tyler, per Beaumont e Wichita Falls il trend è invece abbastanza costante.