Progetto finale del corso in Statistica Descrittiva per Data Scientist di ProfessionAI.

1

```
data <- read.csv("realestate_texas.csv")
```

2

Indica il tipo di variabili contenute nel dataset.

- **city**: var. qualitativa su scala nominale
- **year**: var. quantitativa discreta su scala di intervalli
- **month**: var. quantitativa discreta su scala di intervalli
- **sales**: var. quantitativa discreta su scala di rapporti
- **volume**: var.quantitativa continua su scala di rapporti
- **median_price**: var. quantitativa continua su scala di rapporti
- **listings**: var. quantitativa discreta su scala di rapporti
- **months_inventory**: var. quantitativa continua su scala di rapporti

3

3.1

Per le seguenti variabili non ha senso calcolare indici di posizione, di dispersione e di forma:

- **city**: perché è qualitativa
- **year** e **month**: perché non sono caratteristiche intrinseche delle unità osservate, non le vogliamo studiare di per sè. Servono a definire la granularità temporale delle osservazioni; vale a dire, servono a dividere l'intervallo temporale totale in classi di riferimento con uguale ampiezza, esaustive e mutualmente esclusive.

3.2

Tabella di frequenze per **city** (è equidistribuita):

Beaumont	Bryan-College Station	Tyler	Wichita Falls
60	60	60	60

3.3

Indici di posizione, di dispersione e di forma per le rimanenti variabili:

	Min	1st Qu.	Median	Mean	3rd Qu.	Max	Range	IQR	SD	Var	CV	Asim	Kurtosis
sales	79	127	175.5	192.2	247	423	344	120	79.6	6344.3	41.4	0.72	-0.31
volume	8.1	17.6	27.0	31.0	40.8	83.5	75.3	23.2	16.6	277.2	53.7	0.88	0.18
median_price	73800	117300	134500	132665.4	150050	180000	106200	32750	22662.1	513572983.0	17.0	-0.36	-0.62
listings	743	1026.5	1618.5	1738.0	2056	3296	2553	1029.5	752.7	566568.9	43.3	0.65	-0.79
months_inventory	3.4	7.8	8.9	9.1	10.9	14.9	11.5	3.1	2.3	5.3	25.1	0.04	-0.17

Sono stati calcolati nel modo seguente:

```
statistical_indexes <- function(x){

    position <- summary(x)

    dispersion <- c("Intervallo:" = max(x)-min(x),
                    "IQR:" = IQR(x),
                    "dev.st:" = sd(x),
                    "var:" = var(x),
                    "CV:" = sd(x)/mean(x) * 100)

    shape <- c("Asimmetria" = skewness(x),
               "Curtosi" = kurtosis(x) - 3)

    indexes <- c(position, dispersion, shape)
    return(indexes)
}
```

Osservazioni:

- L'unico valore negativo di asimmetria si ha per median_price. Infatti la sua media semplice è inferiore della sua mediana, quindi sono più frequenti valori alti che valori bassi - a differenze delle altre variabili quantitative, che presentano una distribuzione asimmetrica positiva (valori bassi più frequenti che valori alti).
- months_inventory è la variabile con distribuzione più simmetrica
- Tutte le variabili tranne volume hanno indice di curtosi negativo, quindi seguono una distribuzione platicurtica (più piatta della distribuzione normale). volume segue una distribuzione leptocurtica (più appuntita della normale).

4

La variabile con variabilità più elevata è quella con CV maggiore, quindi volume (CV=53.7).
La variabile più asimmetrica è quella con indice di asimmetria (in valore assoluto) maggiore, quindi volume con 0.88.

5

Divido la variabile `volume` in classi di ampiezza 5 (cinque milioni), e ne calcolo la tabella delle frequenze e l'indice di Gini.

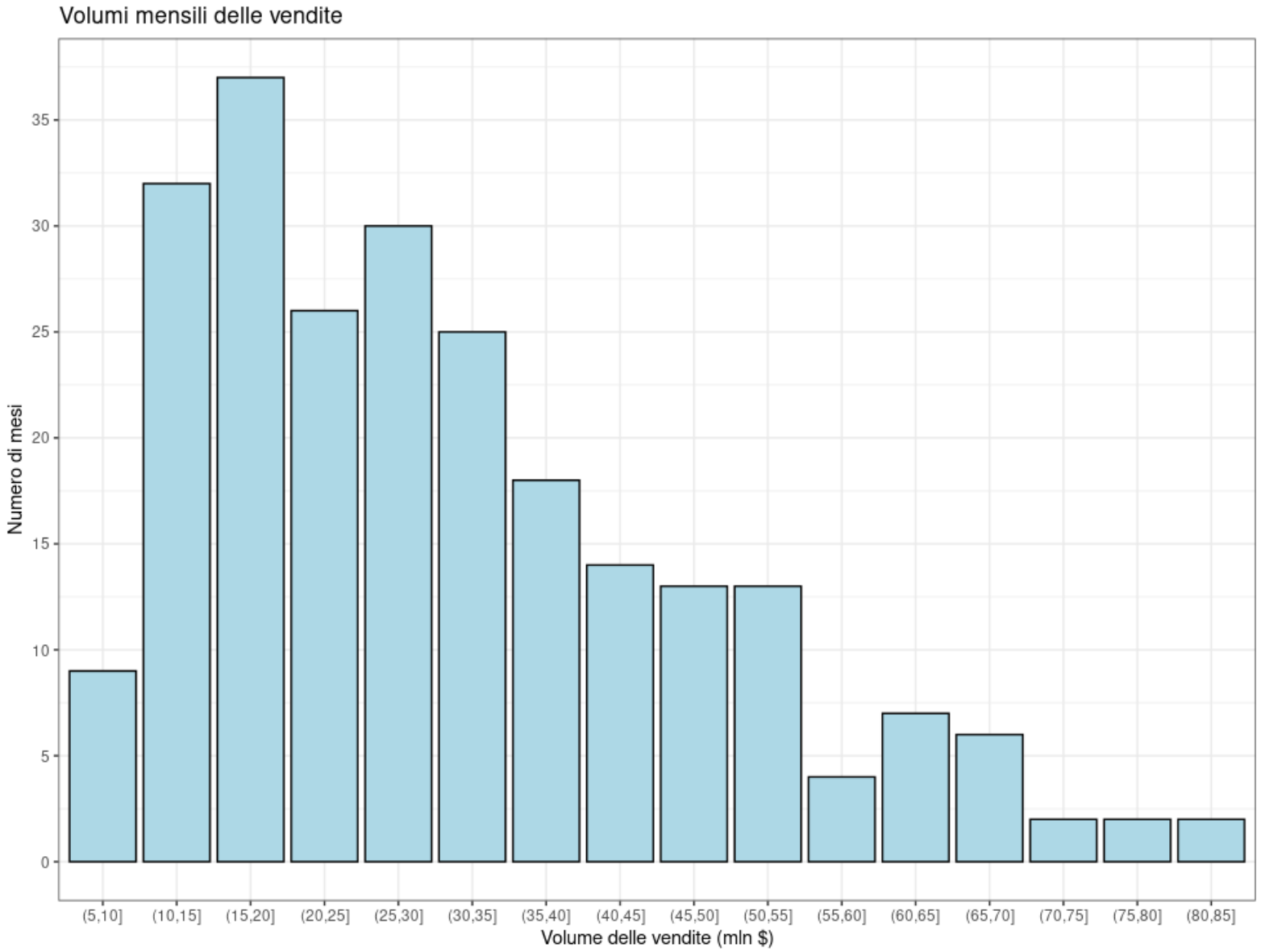
```
volume_cl <- cut(volume, breaks = seq(5,85,5))
freq_table_volume_cl <- table(volume_cl)
```

Tabella delle frequenze:

(5,10]	(10,15]	(15,20]	(20,25]	(25,30]	(30,35]	(35,40]	(40,45]	(45,50]	(50,55]	(55,60]	(55,60]	(65,70]	(70,75]	(75,80]	(80,85]
9	32	37	26	30	25	18	14	13	13	4	7	6	2	2	2

Indice di Gini = 0.962
(infatti vengono assunte tutte le modalità, e non ce n'è una che spicca sulle altre)

Grafico a barre:



6

Le città hanno la stessa frequenza quindi l'indice di eterogeneità di Gini ha valore massimo, cioè 1.

7

```
N <- nrow(data)
p1 <- length(city[city=="Beaumont"]) / N
p2 <- length(month[month==7]) / N
p3 <- nrow(data[data$month==12 & data$year==2012,]) / N
print(c(p1, p2, p3))
```

Le probabilità risultanti sono: 0.25, 0.083 (1/12), 0.016 (1/60)

8

Prezzo medio (media aritmetica):

```
data$mean_price <- (volume / sales) * 1000000
```

9

Un annuncio è efficace se il relativo immobile verrà venduto il prima possibile (idealmente lo stesso mese).
Quindi l'efficacia degli annunci va di pari passo con il numero di vendite mensili fratto il numero di annunci attivi.

(L'efficacia cresce al crescere delle vendite fissato il numero di annunci, e cala al crescere del numero di annunci fissato il numero di vendite).

```
data$listings_effectiveness <- sales / listings
```

10

Manipolazione dati con `dplyr`:

10.1

```
# Average number of sales each month
data %>%
    group_by(mese = month) %>%
    summarise(numero_medio_vendite = mean(sales)) %>%
    pivot_wider(names_from = mese, values_from = numero_medio_vendite)
```

Mese	1	2	3	4	5	6	7	8	9	10	11	12
Numero medio vendite	127	141	189	212	239	244	236	231	182	180	157	169

10.2

```
# Top 2 months by number of sales for each city
data %>%
    group_by(city) %>%
    top_n(sales, n = 2) %>%
    select(citta = city, year, month, num_vendite = sales)
```

Città	Anno	Mese	Numero Vendite
Beaumont	2013	8	273
Beaumont	2014	8	262
Bryan-College Station	2013	7	402
Bryan-College Station	2013	7	403
Tyler	2014	5	388
Tyler	2014	6	423
Wichita Falls	2010	4	167
Wichita Falls	2010	5	165

10.3

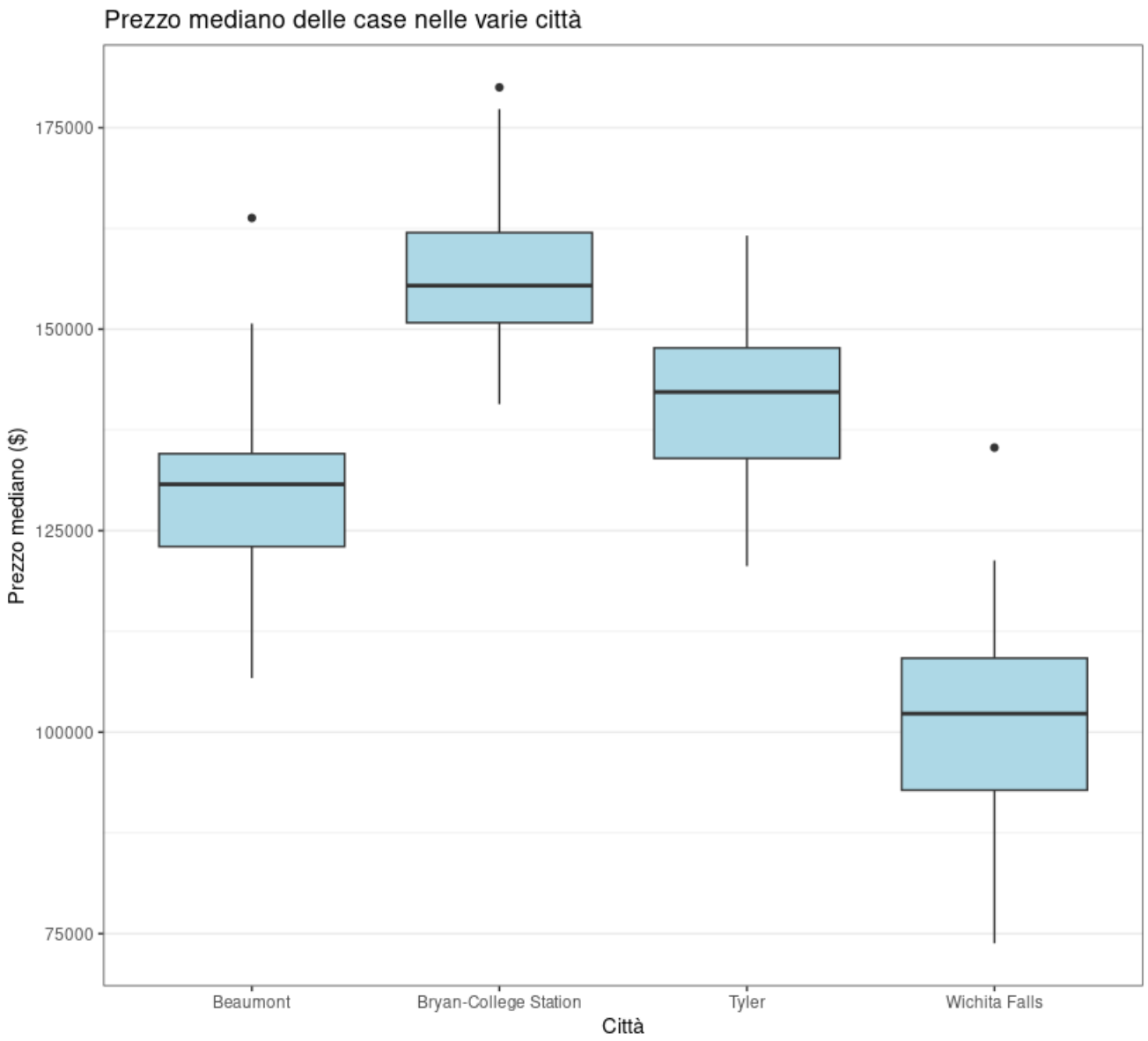
```
# Average number of sales each year in Tyler
data %>%
    filter(city == "Tyler") %>%
    group_by(year) %>%
    summarise(numero_medio_vendite=mean(sales), dev_st=sd(sales))
```

Anno	Numero medio vendite	Dev.st
2010	228	49.0
2011	239	49.6
2012	264	46.4

Anno	Numero medio vendite	Dev.st
2013	287	53.0
2014	332	56.9

11

Confrontando la distribuzione del prezzo mensile mediano delle case vendute tra le varie città, notiamo che la città con i prezzi maggiori è Bryan-College Station mentre quella con i prezzi minori è Wichita Falls.
I prezzi della prima sono circa del 50% superiori rispetto ai prezzi dell'ultima.

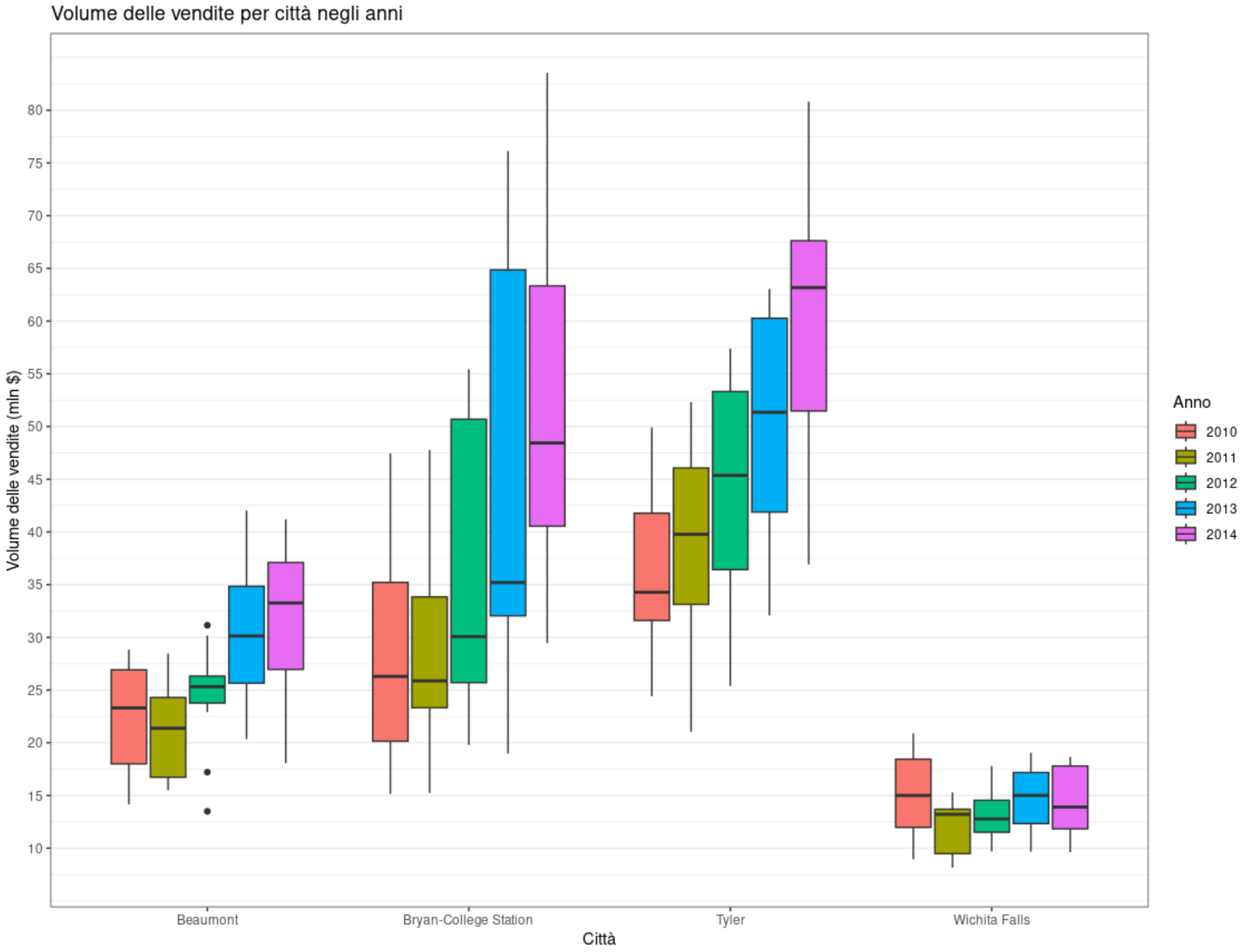



```
ggplot(data=data)+
    geom_boxplot(aes(x = city,
                     y = median_price),
                 fill = "lightblue")+
    labs(title = "Prezzo mediano delle case nelle varie città",
         x = "Città",
         y = "Prezzo mediano ($)")+
```

```
    theme_bw()+
    theme(panel.grid.major.x = element_blank())
```

12

Negli anni il valore totale delle vendite è aumentato in modo marcato ovunque, tranne a Wichita Falls.


Volume delle vendite per città negli anni
Volume delle vendite (mln $)
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
Beaumont
Bryan-College Station
Tyler
Wichita Falls
Città
Anno
2010
2011
2012
2013
2014

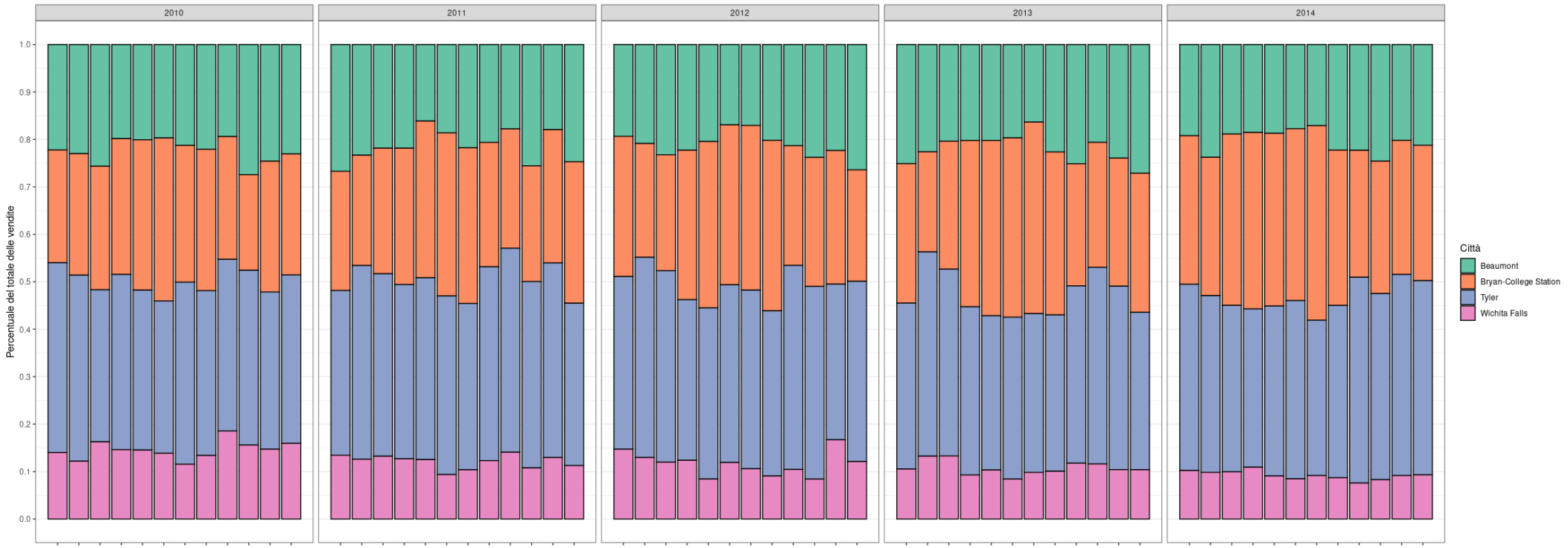
```
ggplot(data=data)+
    geom_boxplot(aes(x = city,
                     y = volume,
                     fill = factor(year)))+
    scale_y_continuous(breaks = seq(10,80,5))+
    labs(title = "Volume delle vendite per città negli anni",
         x = "Città",
         y = "Volume delle vendite (mln $)",
         fill = "Anno")+
    theme_bw()+
    theme(panel.grid.major.x = element_blank())
```

13

- Il valore maggiore delle vendite si osserva nei mesi estivi, in tutti gli anni osservati.
- In percentuale, quanto contribuiscono le singole città a ciò ha una variazione molto piccola negli anni.

Totale delle vendite nei vari mesi

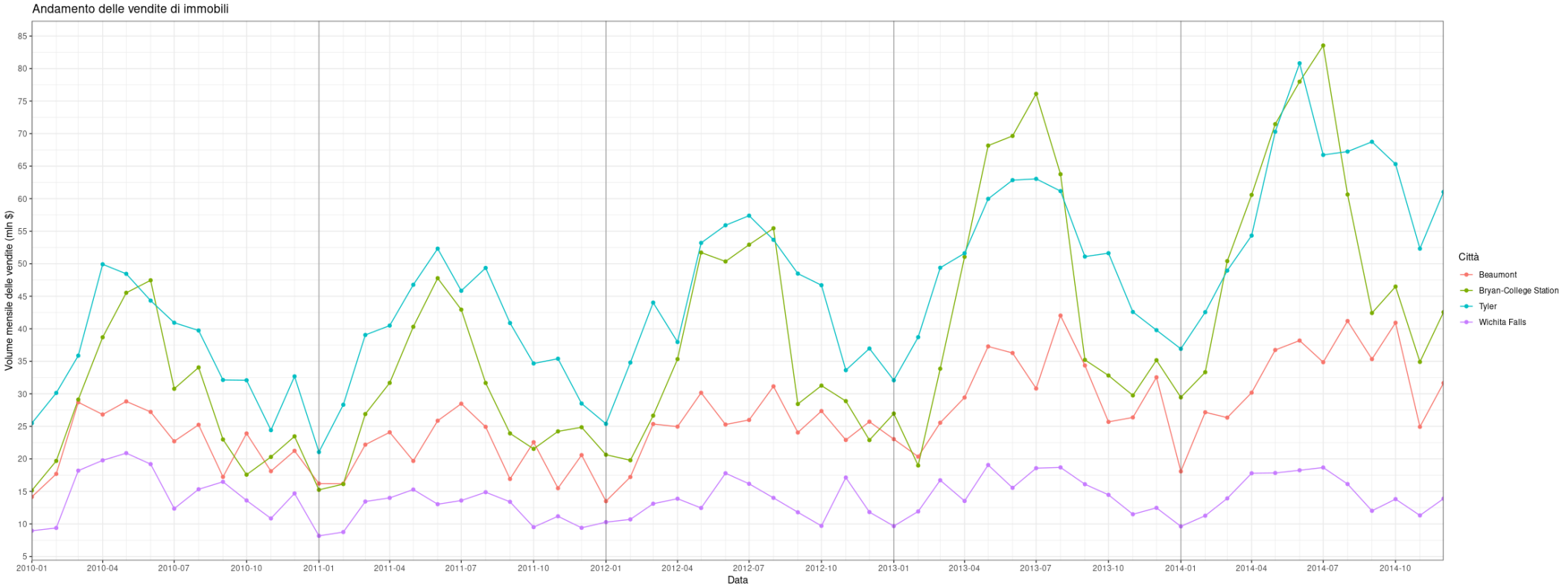
Volume delle vendite per città (%) nei vari mesi




```
# Barre sovrapposte
ggplot(data = data)+
    geom_col(aes(x = month,
                 y = volume,
                 fill = city),
             col = "black")+
    scale_x_continuous(breaks = 1:12)+
    scale_y_continuous(breaks = seq(0,210,10))+
    facet_wrap(~year,
               nrow = 1)+
    scale_fill_brewer(palette = "Set2")+
    labs(title = "Totale delle vendite nei vari mesi",
         x = "Mese",
         y = "Totale delle vendite (mln $)",
         fill = "Città")+
    theme_bw()+
    theme(panel.grid.major.x = element_blank(),
          panel.grid.minor.x = element_blank())


# Barre sovrapposte normalizzate
ggplot(data = data)+
    geom_col(aes(x = month,
                 y = volume,
                 fill = city),
             position = "fill",
             col = "black")+
    scale_x_continuous(breaks = 1:12)+
    scale_y_continuous(breaks = seq(0,1,0.1))+
    facet_wrap(~year,
               nrow = 1)+
    labs(title = "Volume delle vendite per città (%) nei vari mesi",
         x = "Mese",
         y = "Percentuale del totale delle vendite",
         fill = "Città")+
    scale_fill_brewer(palette = "Set2")+
    theme_bw()+
    theme(panel.grid.major.x = element_blank(),
          panel.grid.minor.x = element_blank())
```

- Per tutte le città tranne Wichita Falls, le vendite di immobili sono nettamente cresciute dal 2010 al 2014: per le città di Tyler, Brian-College Station e Beaumont nel 2010-2011 i numeri sono stati simili, nel 2012 il mercato è leggermente cresciuto e nel 2014-2015 è cresciuto ancora di più.
- Generalmente il picco di vendite è sempre d'estate, mentre d'inverno ci sono meno vendite.
- Le città di Tyler e Brian-College Station hanno avuto andamenti simili in tutti gli anni studiati.




```
data$date <- make_date(data$year, data$month)

ggplot(data = data)+
    geom_line(aes(x = date,
                  y = volume,
                  col = city))+
    geom_point(aes(x = date,
                   y = volume,
                   col = city))+
    scale_x_date(date_breaks = "3 months",
                 date_minor_breaks = "month",
                 date_labels = "%Y-%m",
                 expand = c(0, 0))+
    geom_vline(xintercept = date_breaks("1 year")(range(data$date)),
```

```
            lwd = 0.5,
            alpha = 0.3)+
  scale_y_continuous(breaks = seq(5,85,5))+
  labs(title = "Andamento delle vendite di immobili",
       x = "Data",
       y = "Volume mensile delle vendite (mln $)",
       col = "Città")+
  theme_bw()
```