Progetto Mod B

Librerie utilizzate:

Instanziazione del data frame e fattorizzazione delle variabili Region e Channel coi relativi livelli:

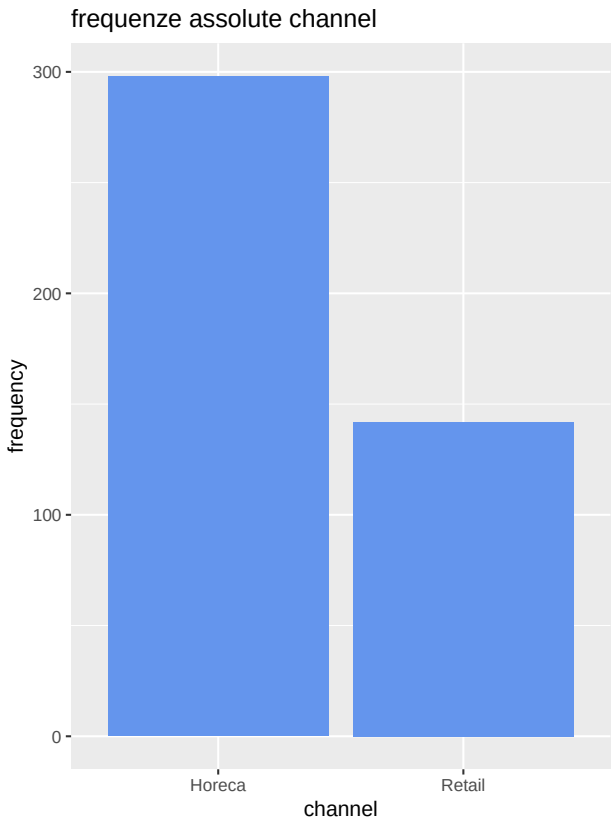
```
[1] "Livelli di Channel:"
```

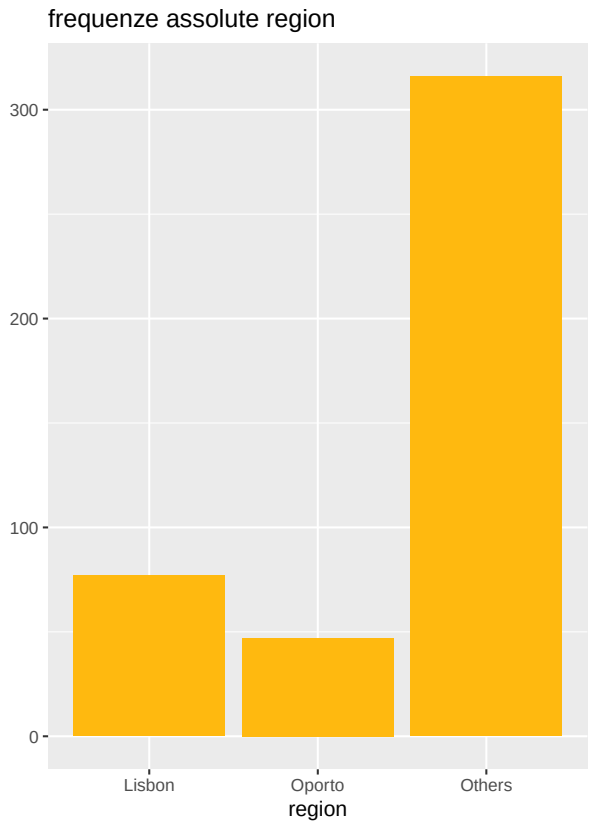
```
[1] "Horeca" "Retail"
```

```
[1] "Livelli di Region:"
```

```
[1] "Lisbon" "Oporto" "Others"
```

Procediamo dunque con una prima analisi esplorativa studiando le frequenze assolute e relative delle due variabili categoriali Channel e Region:






```
[1] "frequenze relative di channel"
```

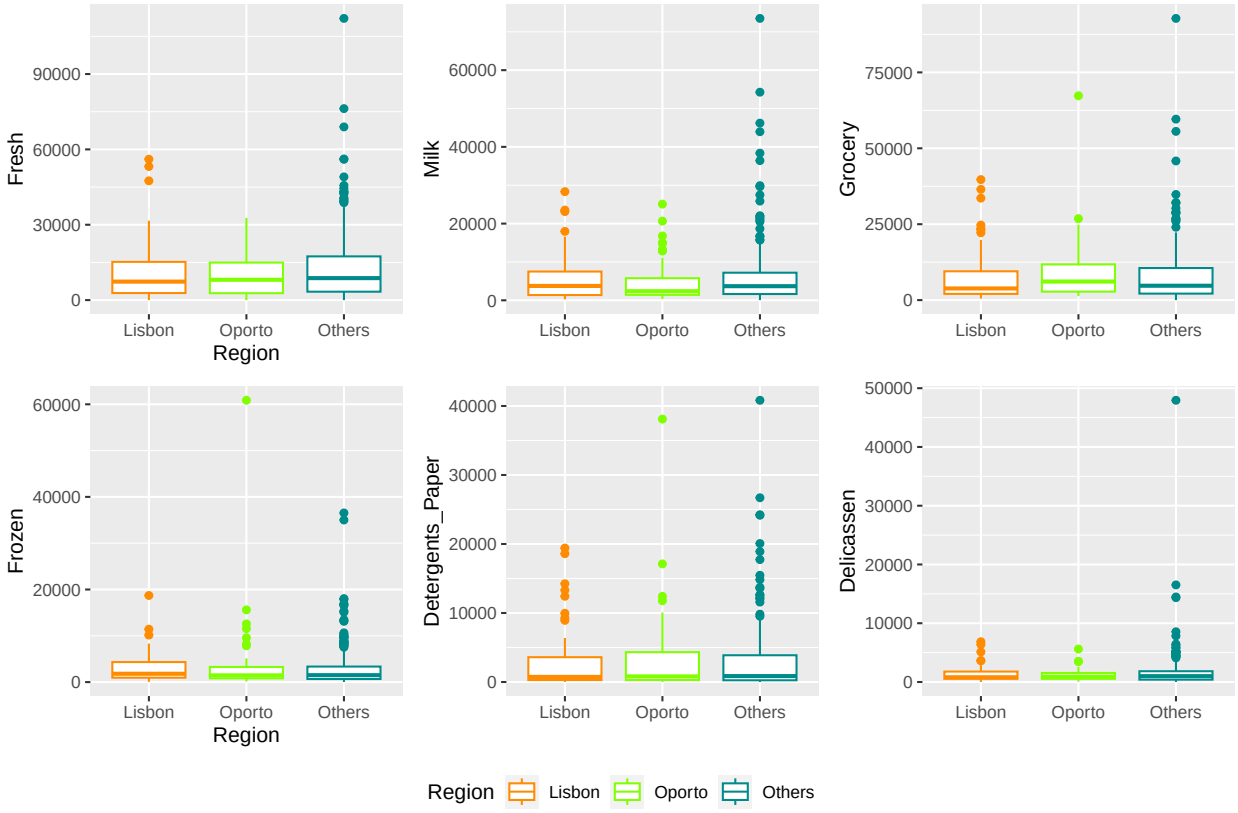
```
Horeca Retail
   298    142
```

```
[1] "frequenze relative di region"
```

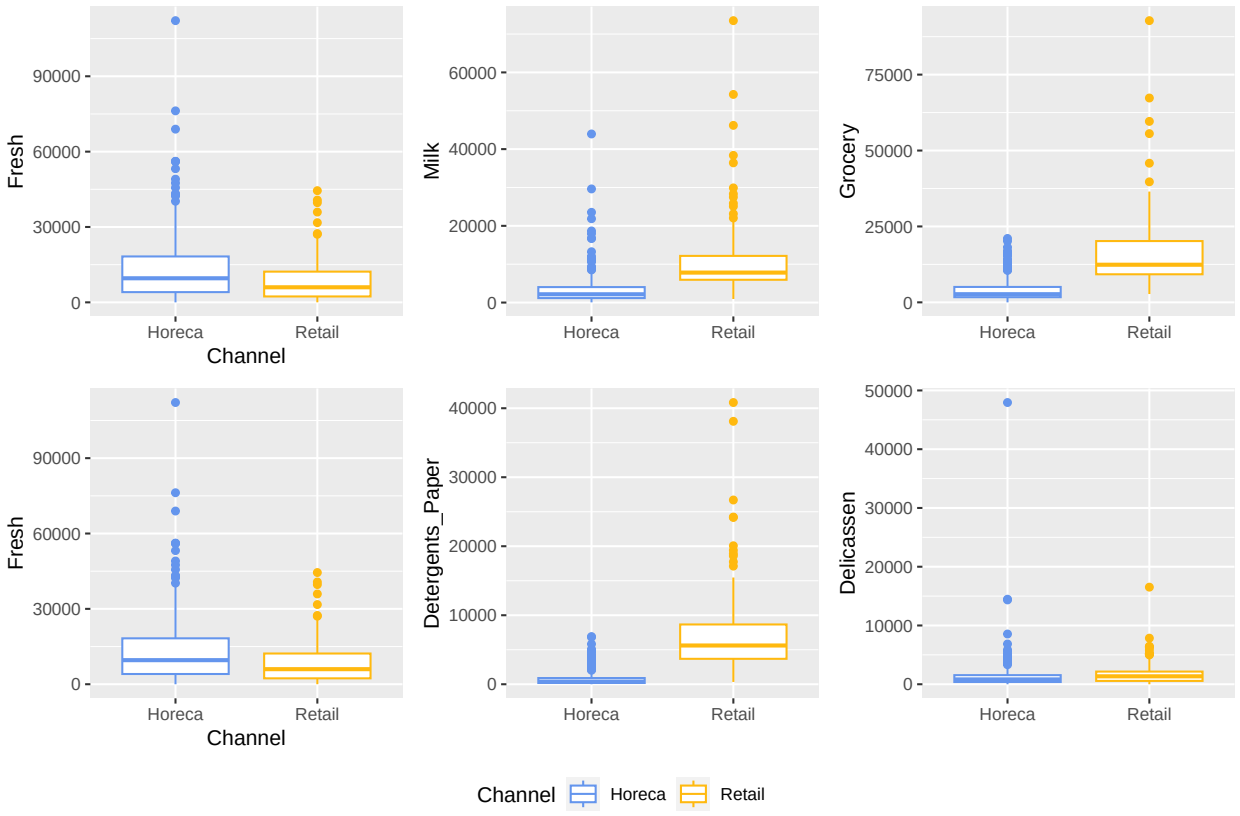
```
Lisbon Oporto Others
    77     47    316
```

Come si può notare dai grafici e dalle tabelle di frequenza, gli ordini piazzati dalla categoria Horeca è decisamente più considerevole di retail (più del doppio). Analogamente la categoria Others presenta decisamente molti più valori rispetto a Lisbon e Oporto.

Passando alle variabili quantitative, possiamo considerare quest'ultime condizionatamente alle due variabili qualitative (Region e Channel) per esplorare eventuali relazioni. Partiamo quindi con la variabile Region:



Procediamo quindi con l'analisi passando alla variabile qualitativa Channel:



Per studiare un'eventuale relazione di dipendenza tra le variabili quantitative, trasformiamo quest'ultime in fattori utilizzando 3 livelli: low, medium e high.

I range considerati sono i seguenti:

- FreshClass:
  - Low: 0-4000
  - Medium: 4001-10000
  - High: 10001-inf
- MilkClass:
  - Low: 0-2000
  - Medium: 2001-6000
  - High: 6001-inf
- GroceryClass:
  - Low: 0-2500
  - Medium: 2501-6000
  - High: 6001-inf
- FrozenClass:

  - Low: 0-4000
  - Medium: 4001-10000
  - High: 10001-inf

- Detergents_PaperClass:
  - Low: 0-500
  - Medium: 501-3000
  - High: 3001-inf

- DelicassenClass:
  - Low: 0-500
  - Medium: 501-1500
  - High: 1501-inf

Possiamo quindi procedere allo studio dell'indipendenza tra le variabili mediante il chi-squared test:

- Channel

```
[1] "p-value di Fresh: 0.008108"

[1] "p-value di Milk: 2.2e-16"

[1] "p-value di Grocery: 2.2e-16"

[1] "p-value di Frozen: 5.836e-06"

[1] "p-value di Detergents_Paper: 2.2e-16"

[1] "p-value di Delicassen: 0.002112"
```

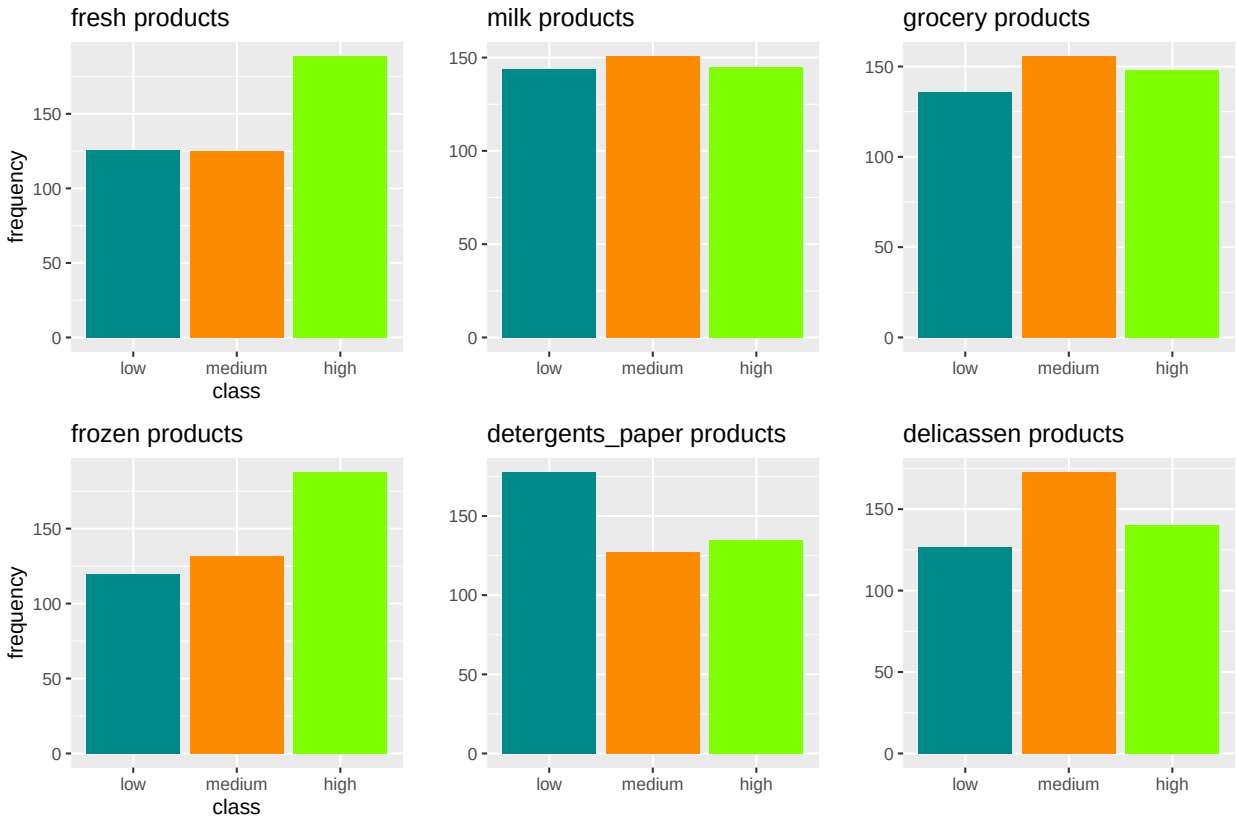
- Region

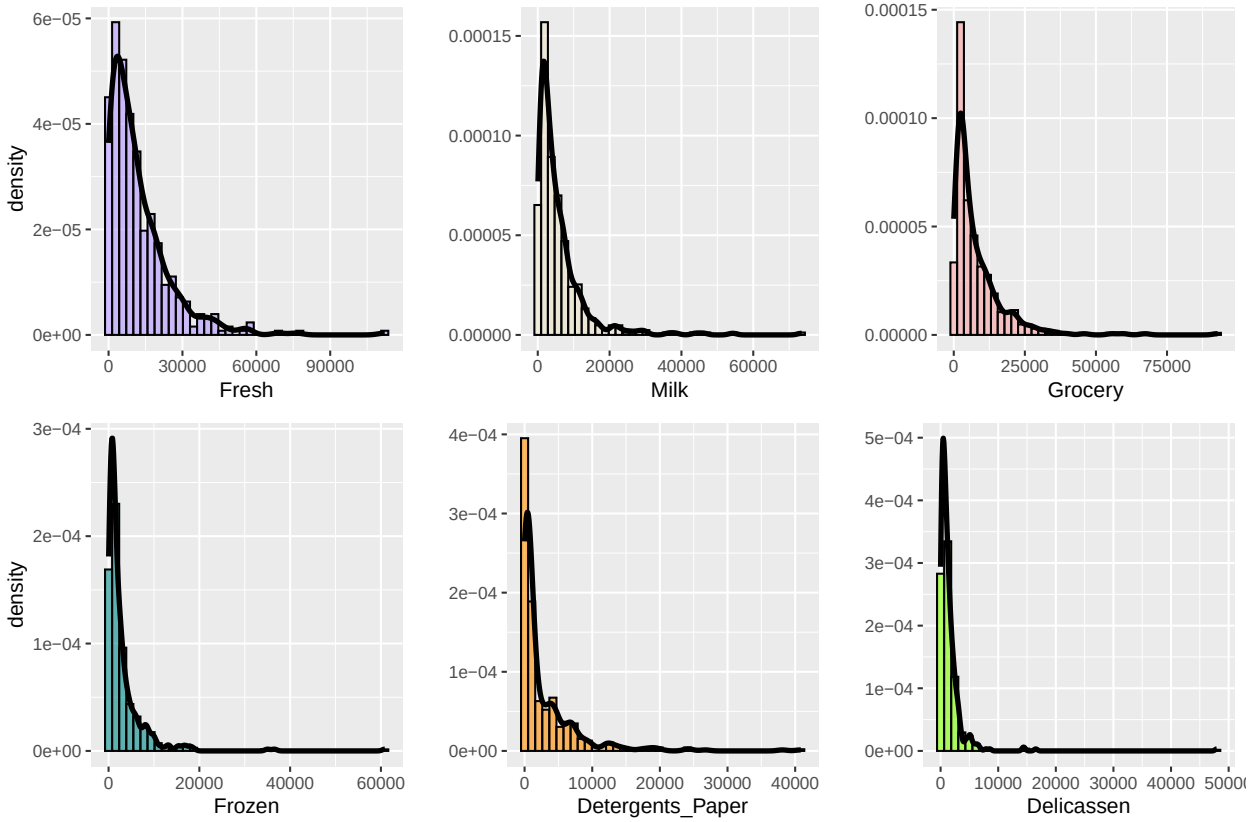
```
[1] "p-value di Fresh: 0.7205"

[1] "p-value di Milk: 0.5231"

[1] "p-value di Grocery: 4224"

[1] "p-value di Frozen: 0.4768"

[1] "p-value di Detergents_Paper: 0.8972"

[1] "p-value di Delicassen: 0.284"
```

Si evince come la variabile Channel influenzi in modo deciso i valori delle spese, in quanto tutte le variabili presentano un p-value inferiore allo 0.05, implicando quindi la presenza di una relazione di dipendenza.

Per quanto riguarda la variabile Region, si nota come i valori del p-value siano superiori alla soglia del 0.05, implicando quindi il rifiuto dell'ipotesi della condizione nulla.

Passiamo ora alla rappresentazione grafica delle variabili quantitative tradotte in classi:




density
6e−05
4e−05
2e−05
0e+00
0
30000
60000
90000
Fresh
0.00015
0.00010
0.00005
0.00000
0
20000
40000
60000
Milk
0.00015
0.00010
0.00005
0.00000
0
25000
50000
75000
Grocery
density
3e−04
2e−04
1e−04
0e+00
0
20000
40000
60000
Frozen
4e−04
3e−04
2e−04
1e−04
0e+00
0
10000
20000
30000
40000
Detergents_Paper
5e−04
4e−04
3e−04
2e−04
1e−04
0e+00
0
10000
20000
30000
40000
50000
Delicassen