Basi di Dati e Data Analytics - Progetto finale

2023-06-28

Vendita all'ingrosso

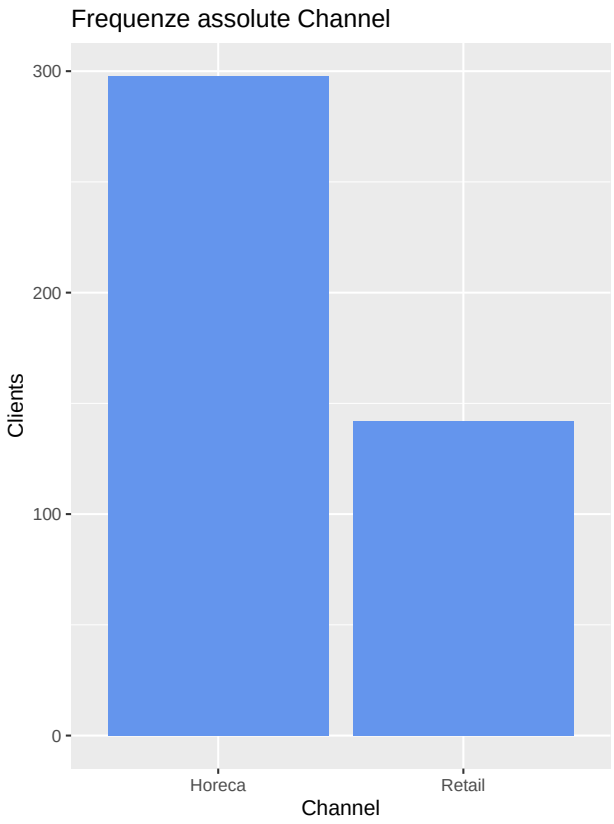
Introduzione

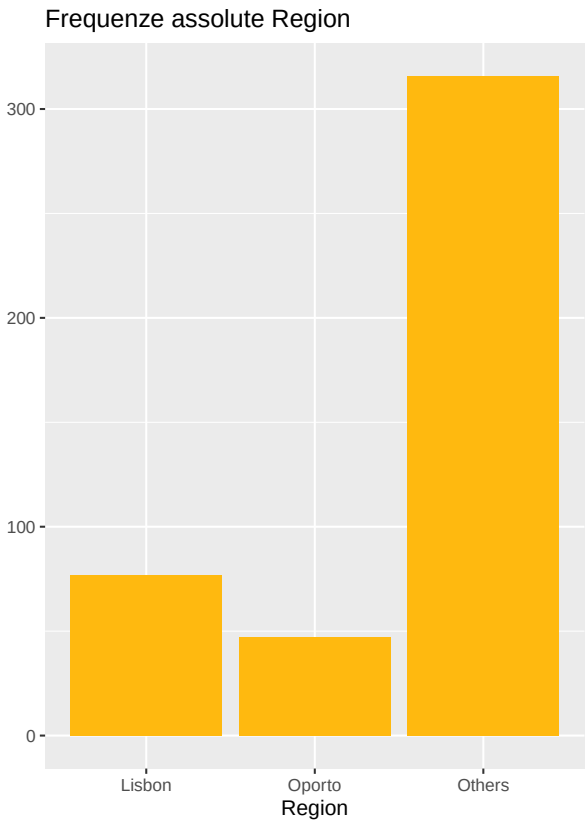
Il dataset considerato in quest'analisi riguarda le spese annuali dei clienti di un commerciante all'ingrosso. I dati sono espressi in unità monetarie ed i clienti vengono raggruppati in due grandi categorie: i luoghi di ristorazione (Horeca) ed i negozi di vendita al dettaglio (Retail). Vengono elencate di seguito le variabili presenti e le relative informazioni:

- Channel [Categoriale], tipologia di venditore (Heroca | Retail);
- Region [Categoriale], regione di provenienza del venditore (Lisbon | Oporto | Other);
- Fresh [Quantitativa], spesa annua di prodotti alimentari freschi;
- Milk [Quantitativa], spesa annua di prodotti alimentari derivanti dal latte;
- Grocery [Quantitativa], spesa annua di prodotti riguardanti drogheria;
- Frozen [Quantitativa], spesa annua di prodotti surgelati;
- Detergents_Paper [Quantitativa], spesa annua di detergenti e prodotti di carta;
- Delicassen [Quantitativa], spesa annua di prodotti di gastronomia.

```
##   Channel Region Fresh Milk Grocery Frozen Detergents_Paper Delicassen
## 1  Retail Others 12669 9656    7561    214             2674       1338
## 2  Retail Others  7057 9810    9568   1762             3293       1776
## 3  Retail Others  6353 8808    7684   2405             3516       7844
## 4  Horeca Others 13265 1196    4221   6404              507       1788
## 5  Retail Others 22615 5410    7198   3915             1777       5185
## 6  Retail Others  9413 8259    5126    666             1795       1451
```

Di seguito riportiamo le rappresentazioni grafiche delle variabili:





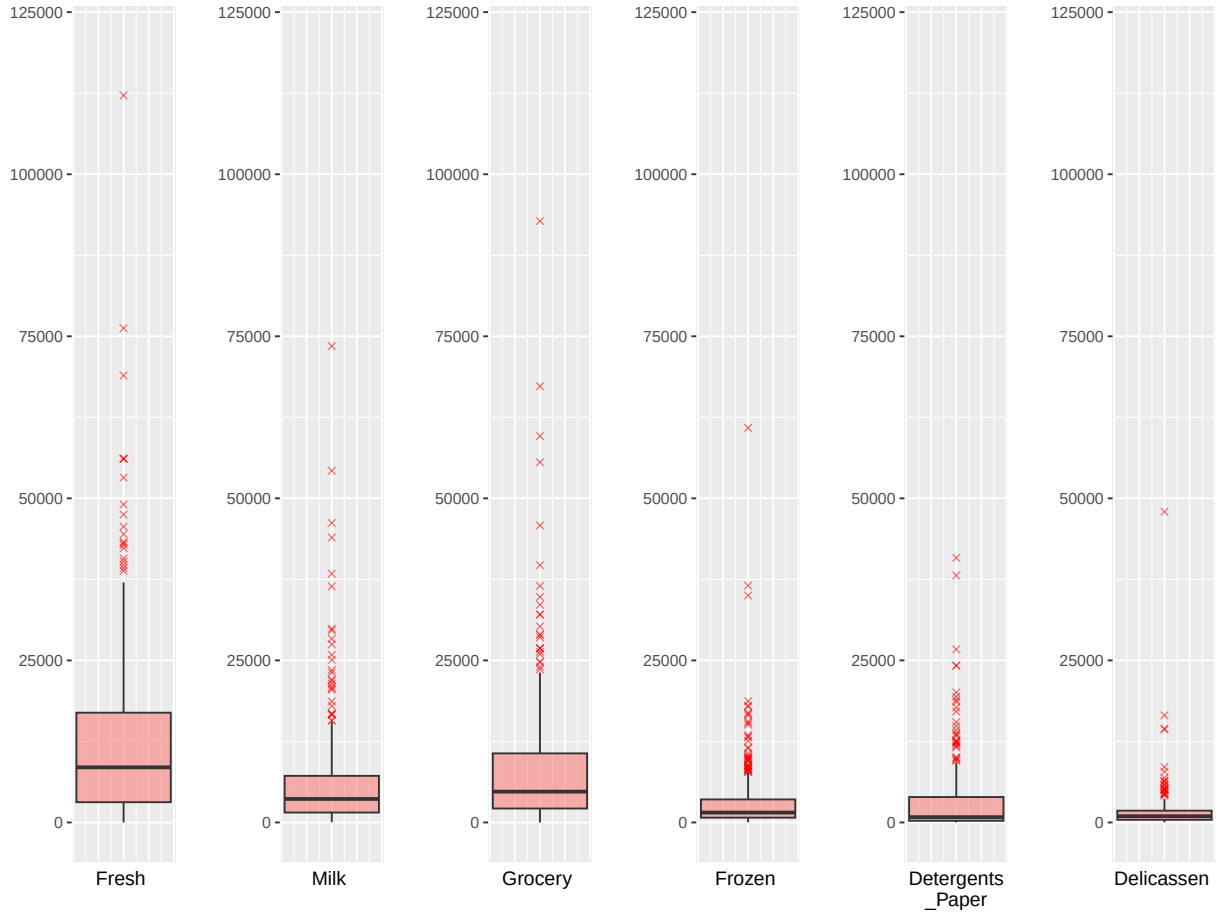

```
## [1] "Frequenze relative di Channel"
```

```
##
## Horeca Retail
##    298    142
```

```
## [1] "Frequenze relative di Region"
```

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

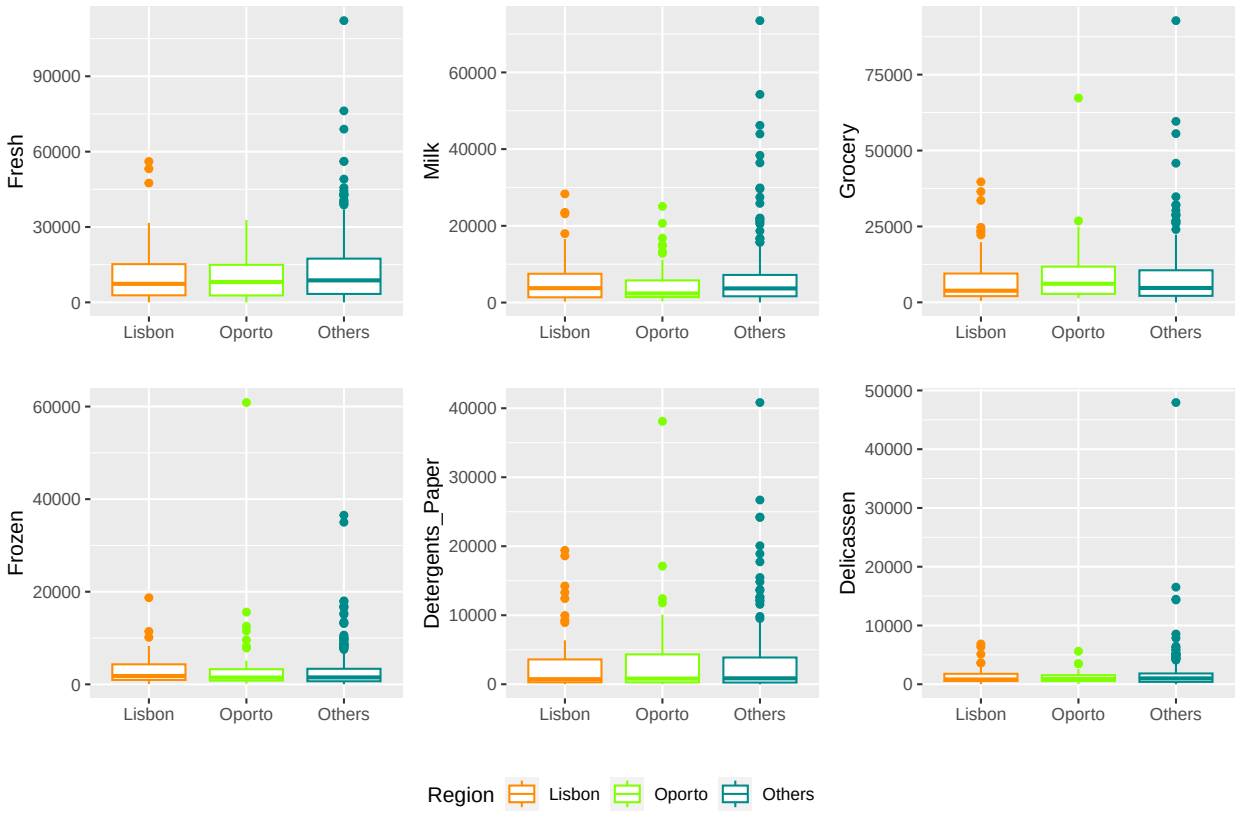
Come si può notare dai grafici e dalle tabelle di frequenza, gli ordini piazzati dalla categoria Horeca sono decisamente più considerevoli di Retail (più del doppio). Analogamente, la categoria Others, presenta decisamente molti più valori rispetto a Lisbon e Oporto.



Dal grafico precedente si nota come le variabili continue presentino distribuzioni differenti, nonostante le relative medie siano pressoché simili. É importante sottolineare la notevole presenza di outliers che potrebbe suggerire un'analisi orientata alla loro eliminazione. Si decide di non procedere in questo senso proprio per la tipologia di dato che esse rappresentano. Essendo rappresentazioni di spese non ci si aspetta che ci siano errori di misurazione (o perlomeno non se ne ha la certezza), quindi per quanto possano essere anomali ed elevati, essi hanno comunque una loro probabilità di realizzazione e quindi risulterebbe fuorviante condurre un'analisi senza di essi.

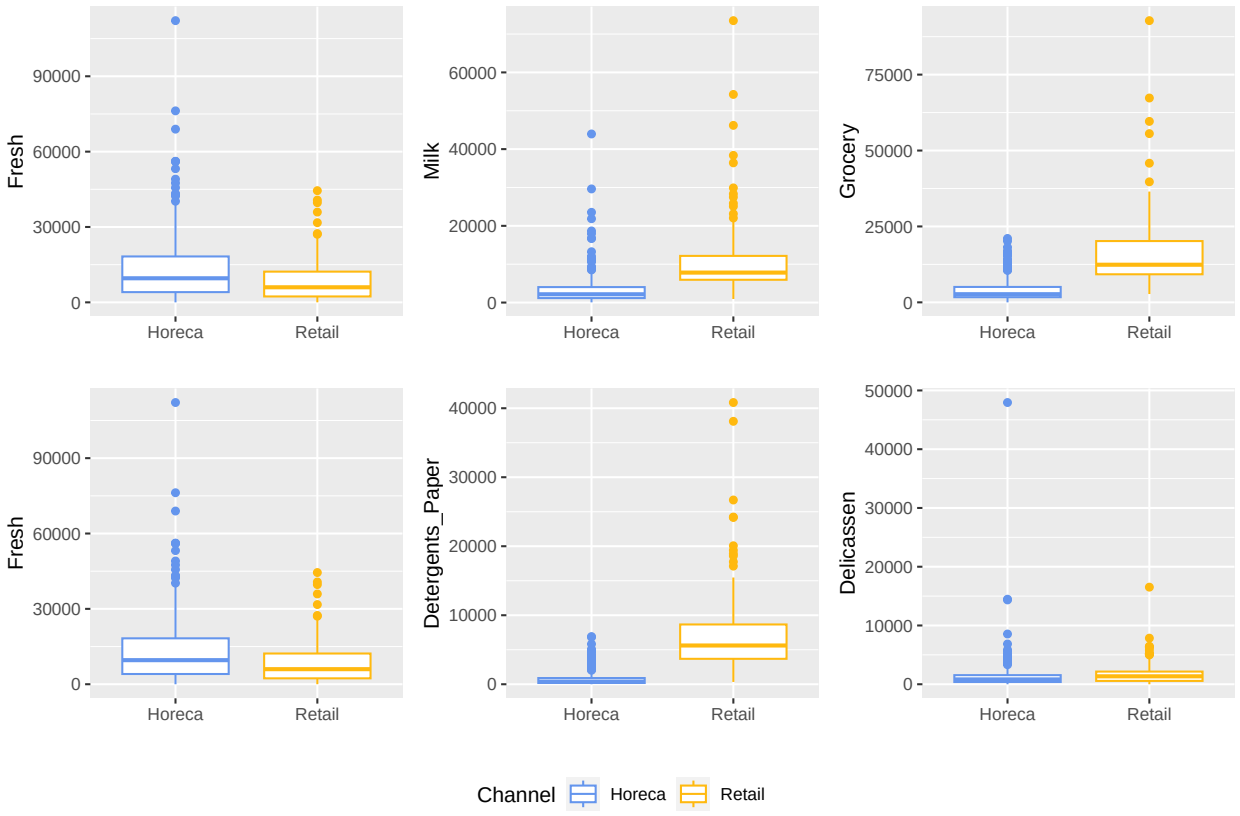
L'obiettivo è quello di individuare le relazioni che intercorrono tra le variabili al fine di generare dei modelli previsionali. Una preliminare fase di analisi esplorativa dei dati può quindi risultare utile.

Consideriamo ora le varibili quantitative condizionatamente alle due variabili qualitative (Region e Channel) per esplorare eventuali relazioni. Partiamo quindi con la variabile Region:



Dai boxplot delle quantitative condizionate alla variabile Region emerge come, quest'ultima, non influenzi le distribuzioni, in quanto i boxplot sono molto simili.

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



Fatta esclusione per Delicassen, ora, le distribuzioni mostrate dai boxplot risultano essere significativamente differenti, sintomo di una sostanziale influenza della variabile 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"
```

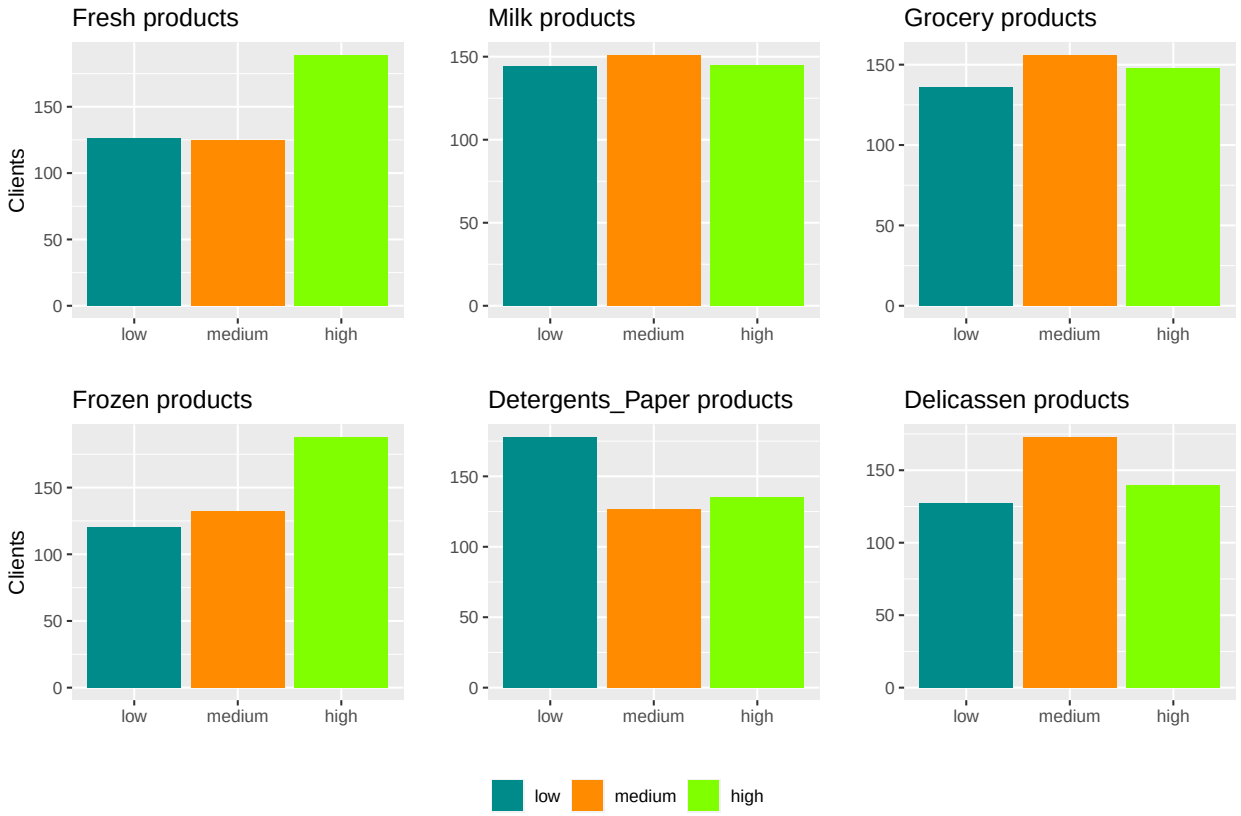
- Region

```
## [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"
```

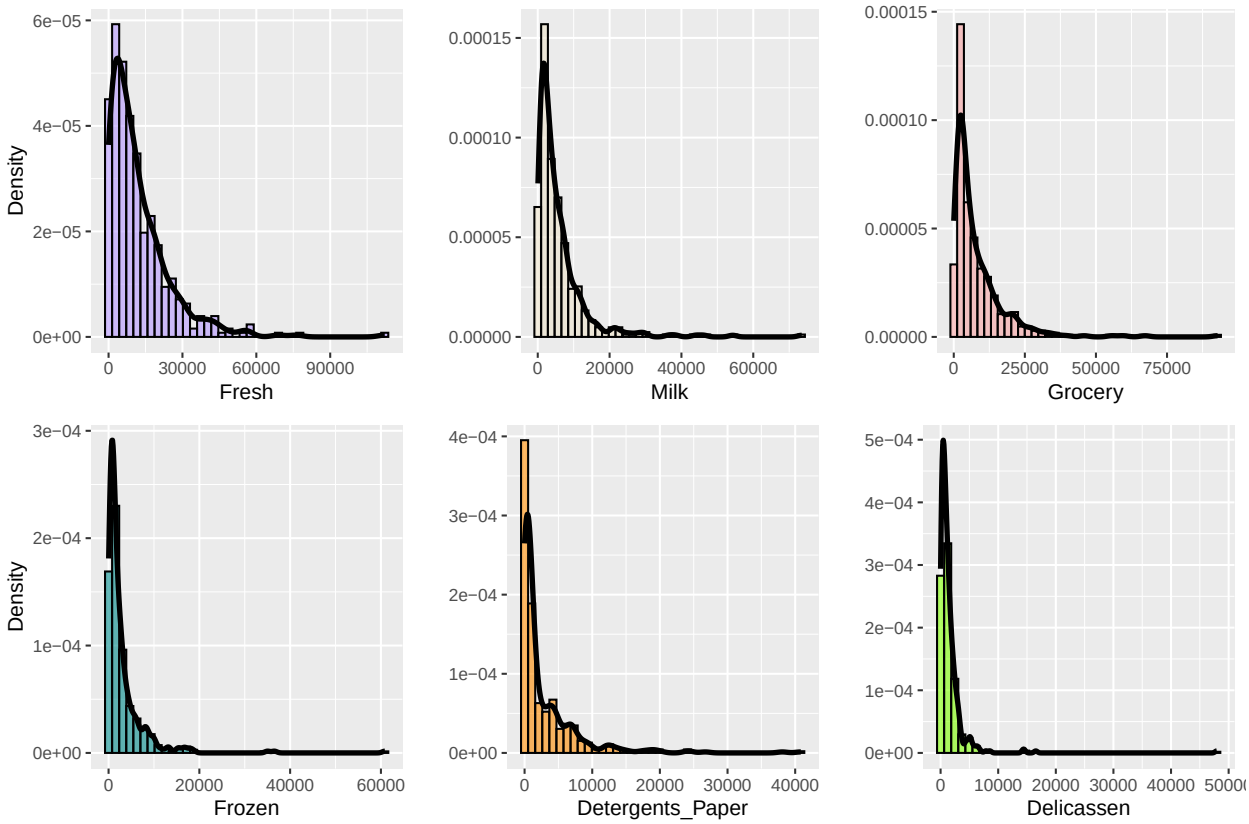
A conferma di quanto detto sopra, si evince come la variabile Channel influenzi in modo deciso i valori delle spese, in quanto tutte le variabili presentano un p-value inferiore a 0.05, implicando quindi la presenza di una relazione di dipendenza.

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

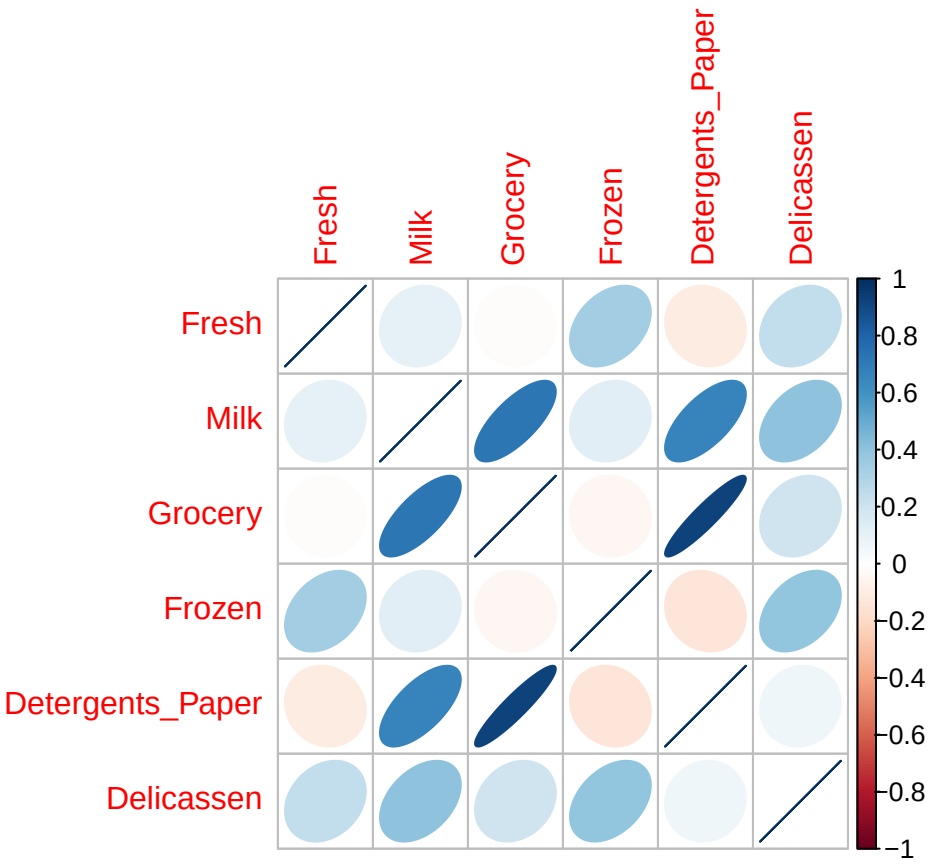
Passiamo ora alla rappresentazione grafica delle variabili quantitative tradotte in classi:

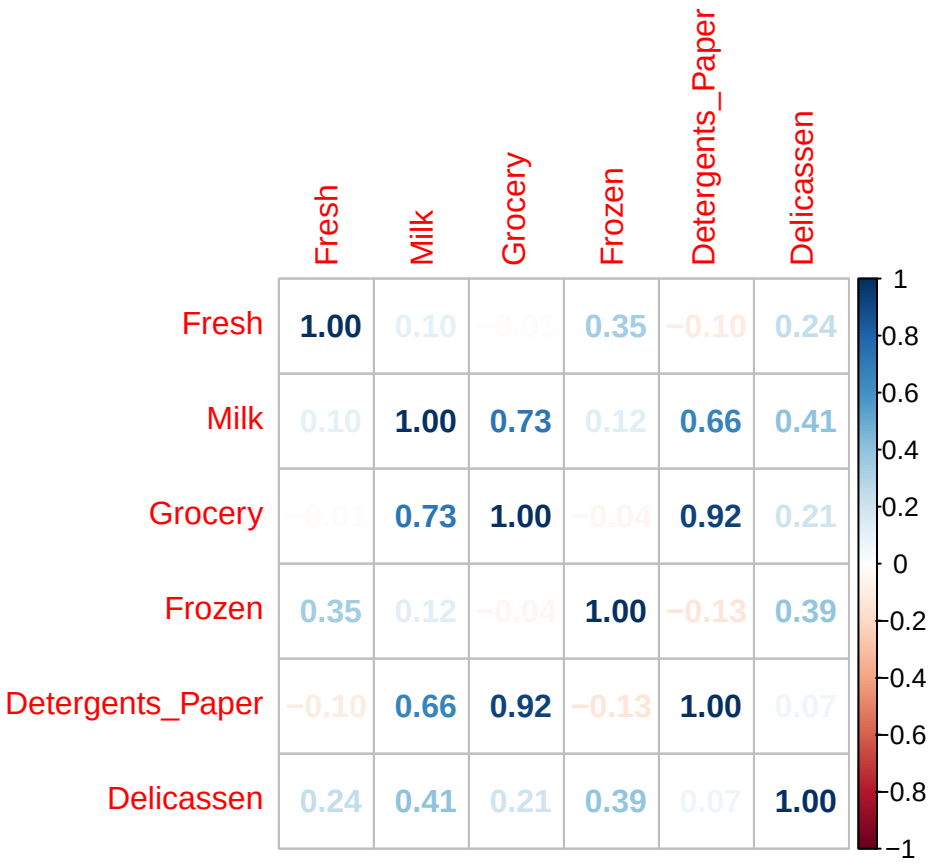


Plottiamo ora gli istogrammi delle varibili quantitative a cui sovrapponiamo le curve di densità.



Ora analizziamo la matrice di correlazione con con il comando corrplot per vedere quali sono gli indici di correlazione tra le variabili presenti nel dataset.


Fresh
Milk
Grocery
Frozen
Detergents_Paper
Delicassen
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1



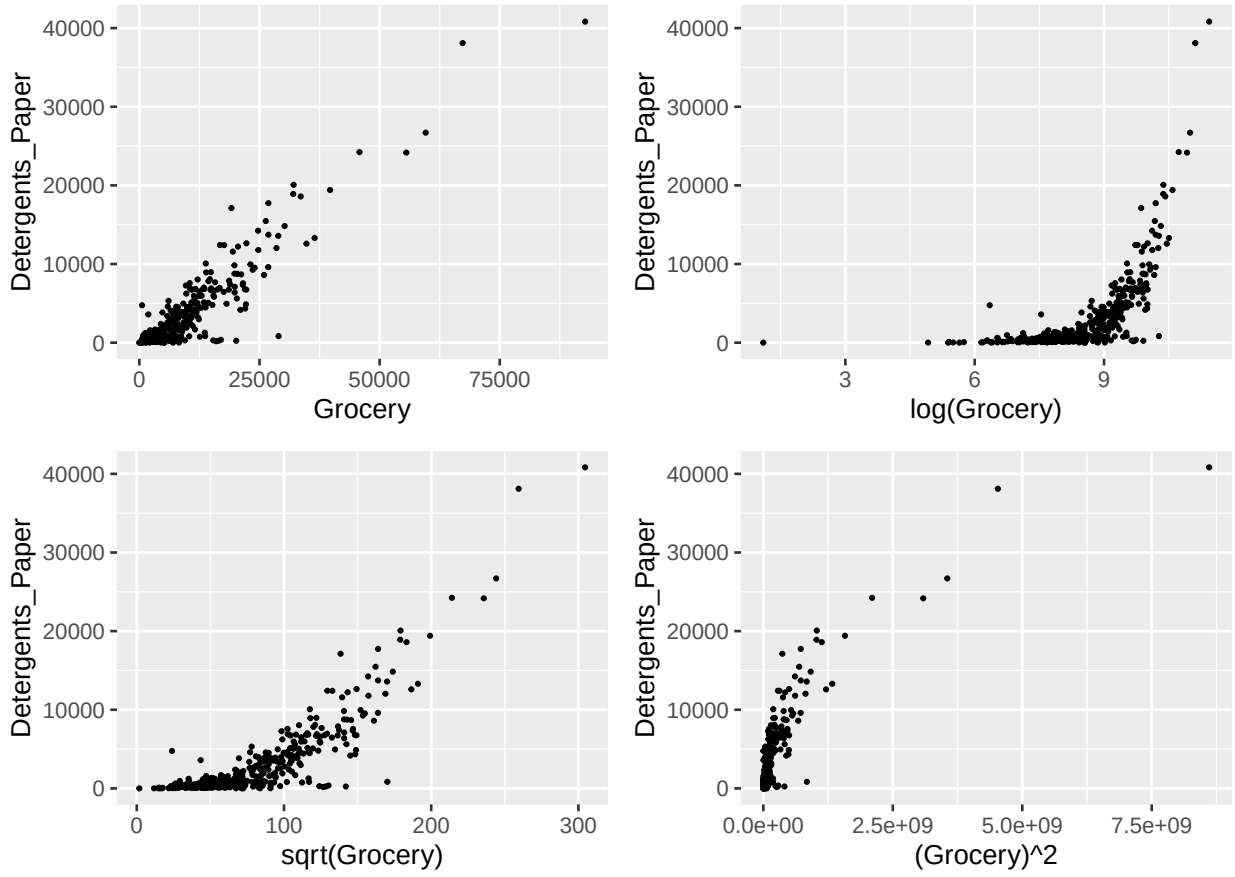
Successivamente all'analisi del corrplot abbiamo selezionato 4 coppie di variabili, le 4 con indice di correlazione maggiore, ovvero:

1. Detergents_Paper ~ Grocery: 0.92
2. Milk ~ Grocery: 0.73
3. Detergents_Paper ~ Milk: 0.66
4. Milk ~ Delicassen : 0.41

In seguito considereremo, nell'analisi, anche la possibilità di utilizzare la variabile esplicativa trasformata, quindi con una delle seguenti: trasformazioni:

1. Logaritmica
2. Square root
3. Quadratica

Partiamo con l'analizzare Detergents_Paper ~ Grocery.



- Indice di correlazione senza trasformazioni: 0.9246407
- Indice di correlazione con trasformazione logaritmica: 0.6667038
- Indice di correlazione con trasformazione square root: 0.8505584
- Indice di correlazione con trasformazione quadratica: 0.8138609

Dopo aver valutato gli indici di correlazione con le diverse trasformazioni abbiamo deciso di utilizzare le variabili senza trasformazioni.

Gli R^2 relativi ai tre modelli sono i seguenti:

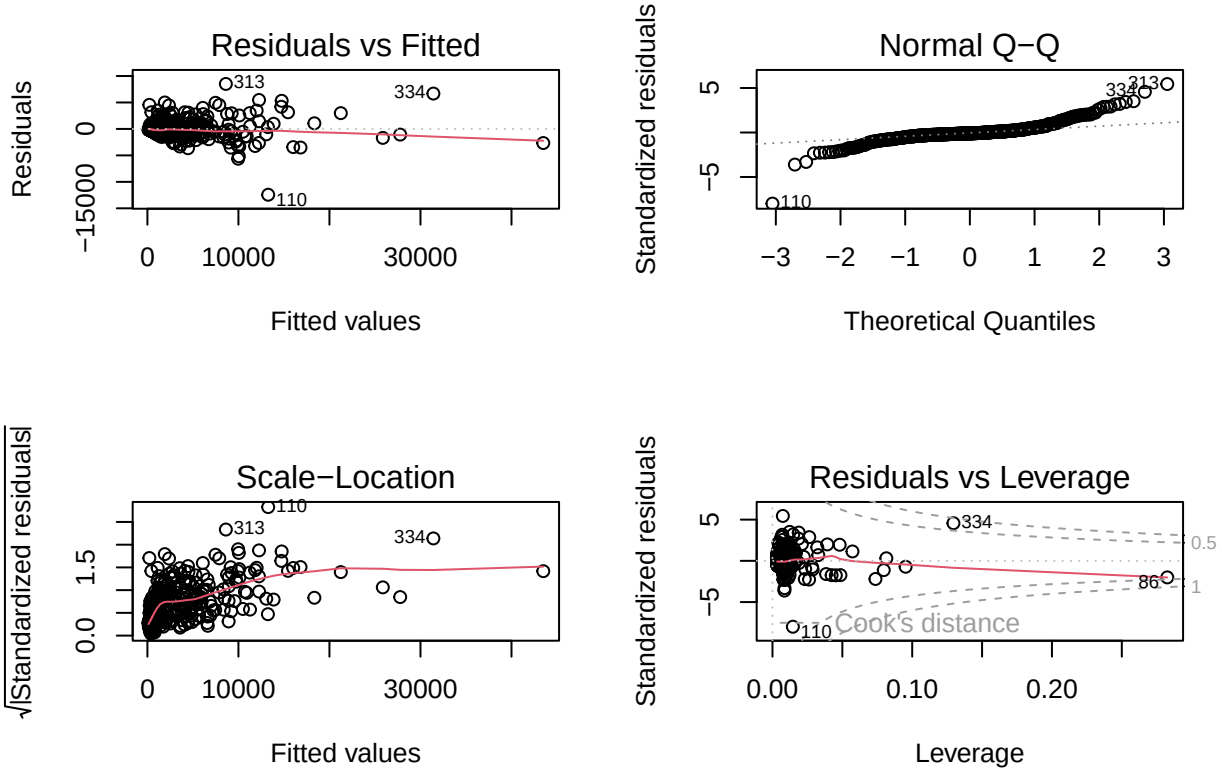
- Semplice: 0.855
- Con aggiunta della variabile Channel: 0.8628
- Con aggiunta dell'interazione con Channel: 0.8919

Output AIC ed ANOVA:

```
##       df      AIC
## fit11  3 7857.429
## fit21  4 7832.878
## fit31  5 7728.892
```

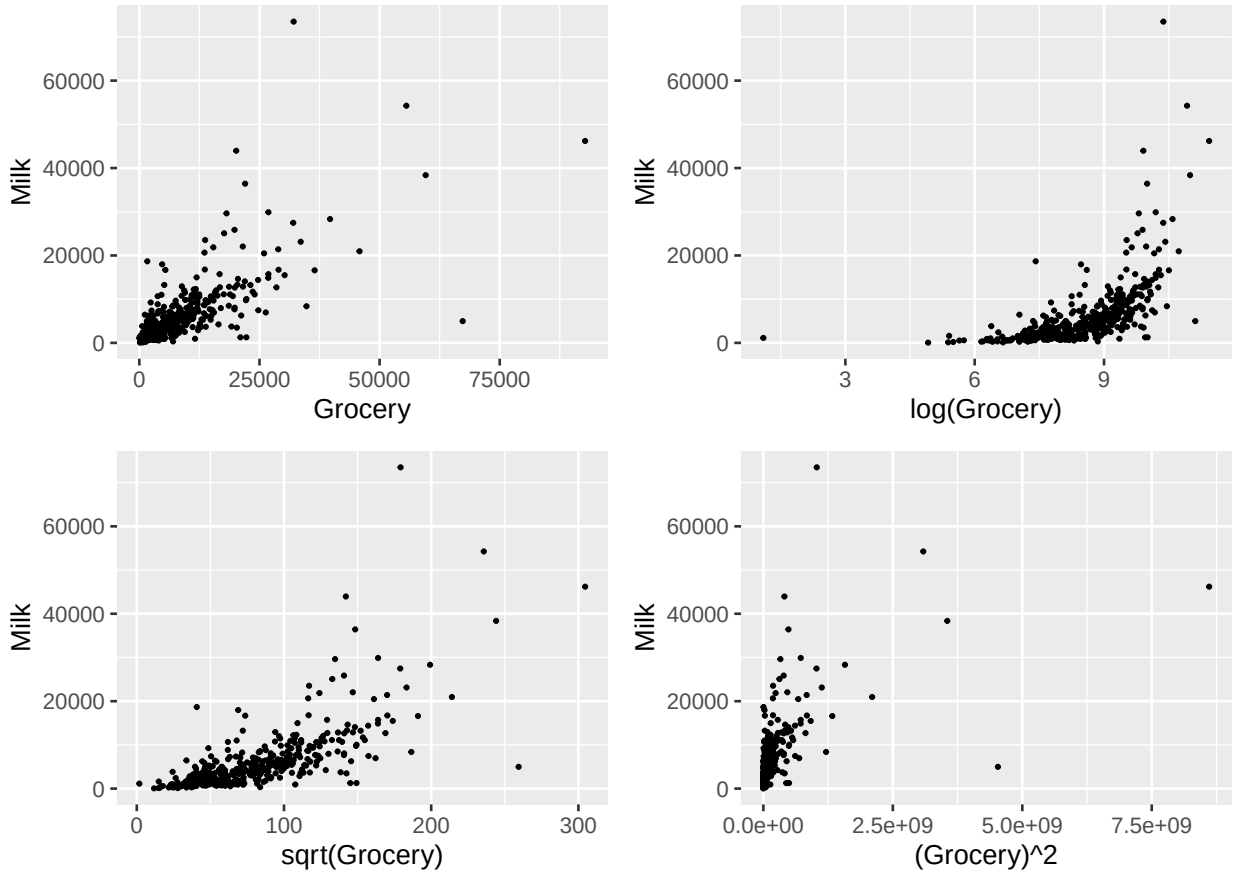
```
## Analysis of Variance Table
##
## Model 1: risp_var1 ~ esp_var1
## Model 2: risp_var1 ~ esp_var1 + Channel
## Model 3: risp_var1 ~ esp_var1 + Channel + esp_var1 * Channel
##   Res.Df        RSS Df Sum of Sq       F    Pr(>F)
## 1    438 1447428334
## 2    437 1362669731  1  84758603  34.506 8.442e-09 ***
## 3    436 1070974396  1 291695335 118.751 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Dopo aver creato i tre modelli, quello semplice, quello con l'aggiunta della variabile categoriale Channel ed, infine, quello con anche l'interazione tra la variabile esplicativa e Channel, abbiamo deciso di utilizzare quest'ultimo, considerando, altresì, i tre relativi AIC ed il test ANOVA.



Dai grafici e dai risultati sulle assunzioni mostrate dall'oggetto gvlma, si può notare come vi sia omoschedasticità ed una discreta gaussianità, tuttavia non sembra esserci una relazione lineare tra le variabili Detergents_Paper e Grocery.

Ora analizziamo Milk ~ Grocery



- Indice di correlazione senza trasformazioni: 0.7283351
- Indice di correlazione con trasformazione logaritmica: 0.5970228
- Indice di correlazione con trasformazione square root: 0.7136706
- Indice di correlazione con trasformazione quadratica: 0.5693098

In seguito alla valutazione degli indici di correlazione con le diverse trasformazioni abbiamo deciso di utilizzare le variabili senza trasformazioni

Gli R^2 relativi ai tre modelli sono i seguenti:

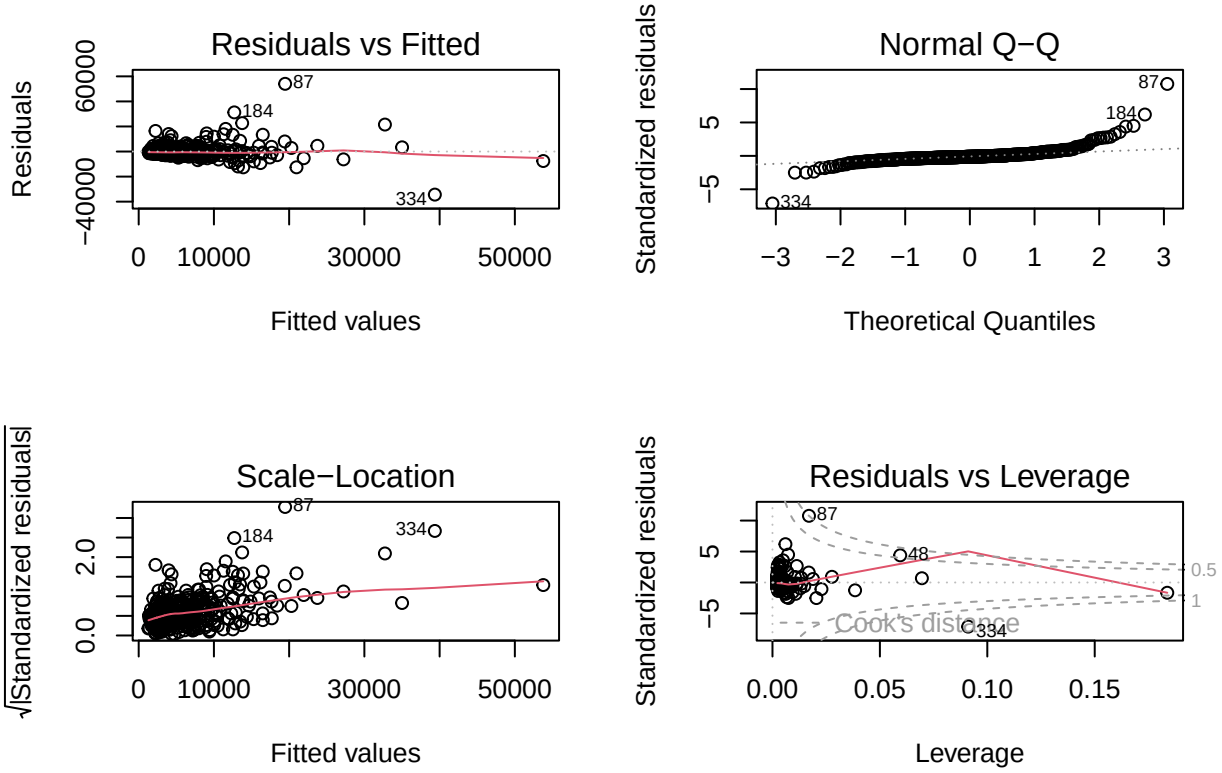
- Semplice: 0.5305
- Con aggiunta della variabile Channel: 0.5288
- Con aggiunta dell'interazione con Channel: 0.5342

Output AIC ed ANOVA:

```
##       df      AIC
## fit12  3 8758.803
## fit22  4 8760.356
## fit32  5 8756.287
```

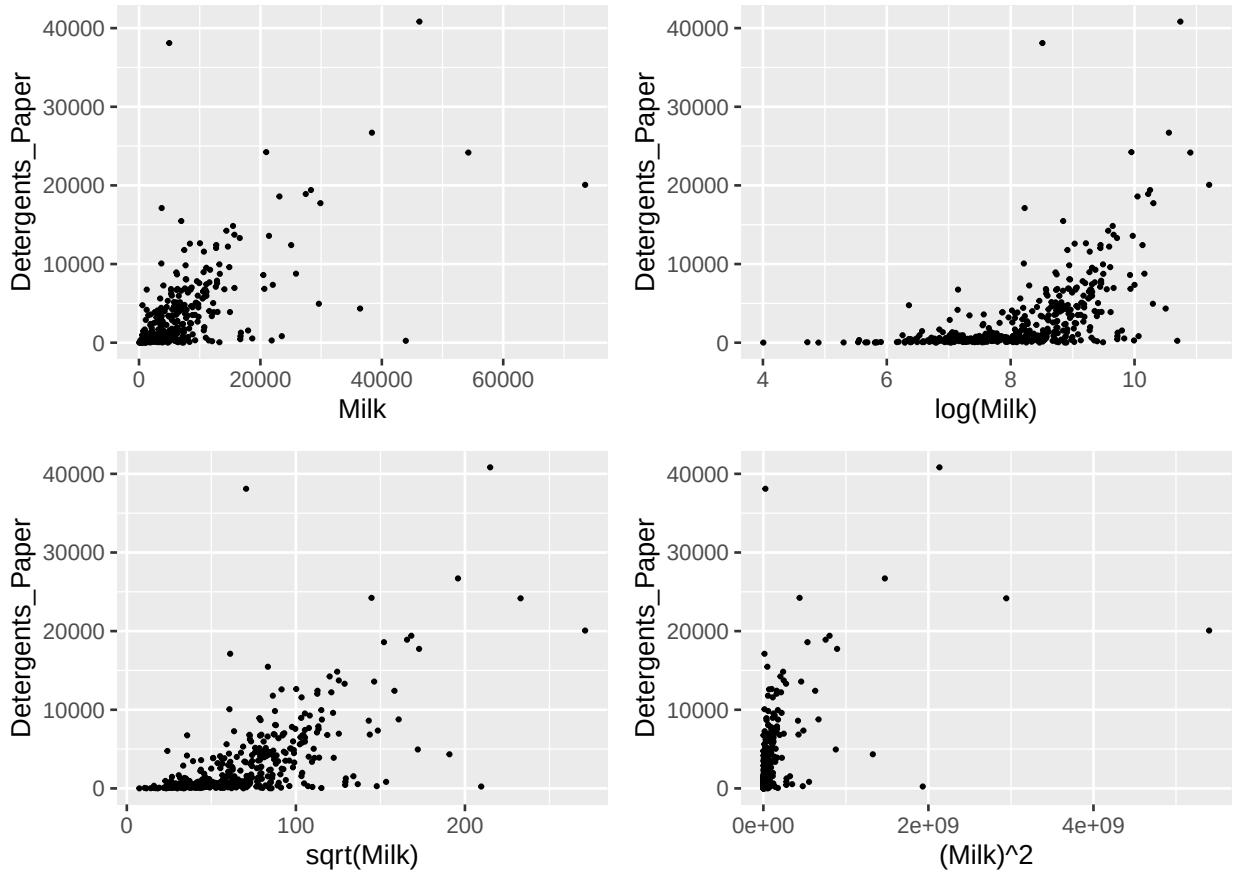
```
## Analysis of Variance Table
##
## Model 1: risp_var2 ~ esp_var2
## Model 2: risp_var2 ~ esp_var2 + Channel
## Model 3: risp_var2 ~ esp_var2 + Channel + esp_var2 * Channel
##   Res.Df        RSS Df Sum of Sq      F  Pr(>F)
## 1    438 1.1228e+10
## 2    437 1.1216e+10  1  11391596 0.4490 0.50318
## 3    436 1.1062e+10  1 153661057 6.0562 0.01424 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Secondo gli stessi criteri di valutazione utilizzati per la coppia di variabili precedente abbiamo deciso di utilizzare il modello semplice, in quanto, nonostante il modello con l'interazione abbia AIC minore e R^2 aggiustato maggiore, la differenza non è così significativa, come emerge dal test ANOVA, pertanto abbiamo preferito utilizzare un modello più semplice possibile.



Dai grafici e dai risultati sulle assunzioni mostrate dall'oggetto gvlma, si può notare come vi sia eteroschedasticità, non gaussianità, e non linearità nell'associazione tra la variabile Milk condizionatamente a Grocery.

Come penultima coppia di variabili analizzate procederemo con Detergents_Paper ~ Milk



- Indice di correlazione senza trasformazioni: 0.6618157
- Indice di correlazione con trasformazione logaritmica: 0.5635915
- Indice di correlazione con trasformazione square root: 0.6578422
- Indice di correlazione con trasformazione quadratica: 0.5154587

In seguito alla valutazione degli indici di correlazione con le diverse trasformazioni abbiamo deciso di utilizzare le variabili senza trasformazioni

Gli R^2 relativi ai tre modelli sono i seguenti:

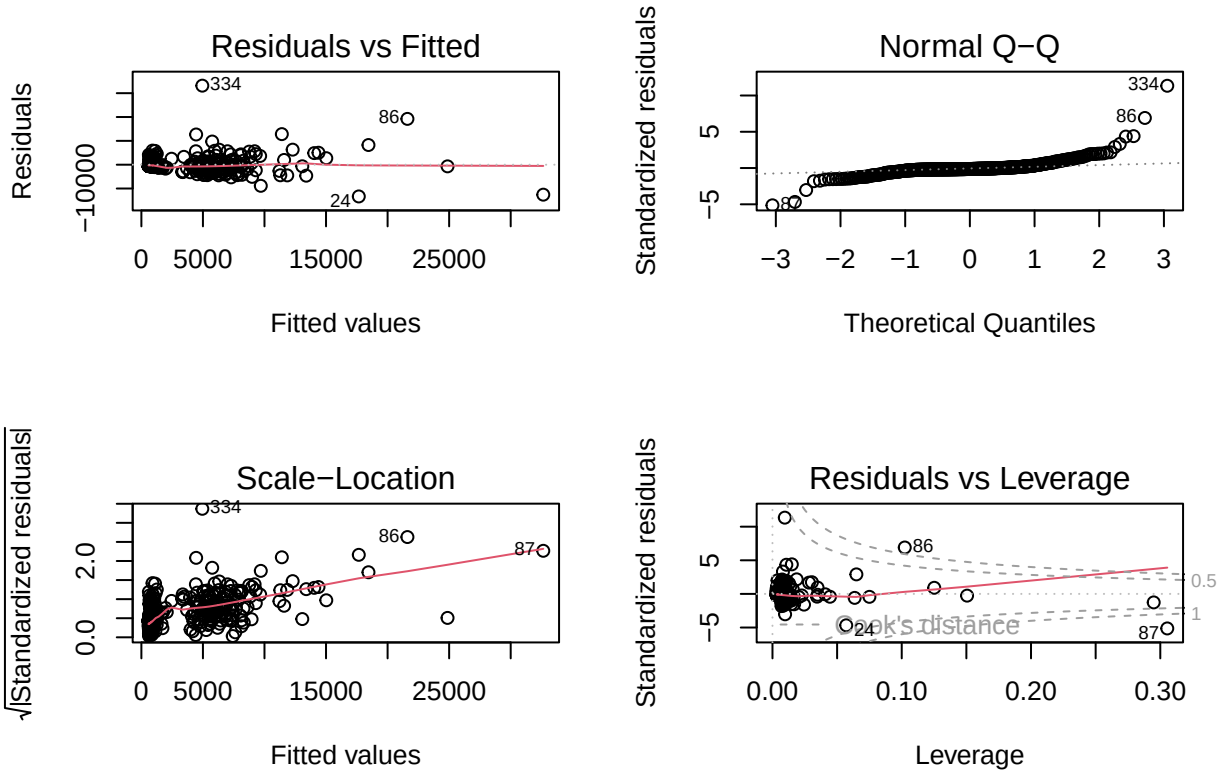
- Semplice: 0.438
- Con aggiunta della variabile Channel: 0.5752
- Con aggiunta dell'interazione con Channel: 0.6205

Output AIC ed ANOVA:

```
##       df      AIC
## fit13  3 8453.407
## fit23  4 8330.203
## fit33  5 8281.613
```

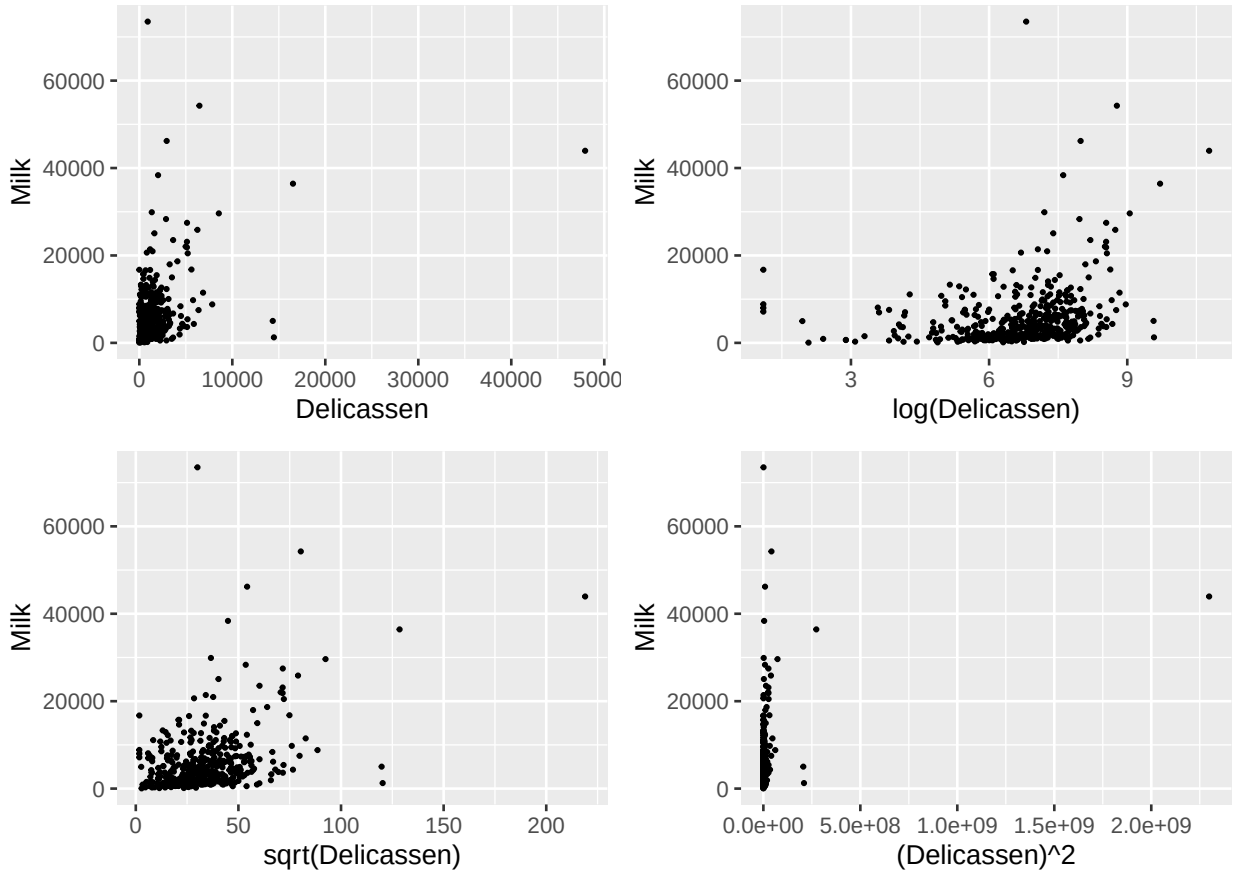
```
## Analysis of Variance Table
##
## Model 1: risp_var3 ~ esp_var3
## Model 2: risp_var3 ~ esp_var3 + Channel
## Model 3: risp_var3 ~ esp_var3 + Channel + esp_var3 * Channel
##   Res.Df        RSS Df  Sum of Sq       F    Pr(>F)
## 1    438 5608501219
## 2    437 4219554287  1 1388946933 161.005 < 2.2e-16 ***
## 3    436 3761256599  1  458297687  53.125 1.481e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Considerando i tre modelli, quello semplice, quello con l'aggiunta della variabile categoriale Channel ed, infine, quello con anche l'interazione tra la variabile esplicativa Milk e Channel, abbiamo deciso di utilizzare quest'ultimo, considerando, altresì, secondo quanto mostrato anche dagli output relativi ai tre AIC ed al test ANOVA



Dai plot sui residui, dal qqnorm e dai risultati delle assunzioni mostrate dall'oggetto gvlma, ne evince come vi sia omoschedasticità, ma per quanto riguarda linearità e gaussianità il modello non rispetta le assunzioni stabilite.

Infine valutiamo come si comporta l'associazione Milk ~ Delicassen



- Indice di correlazione senza trasformazioni: 0.4063683
- Indice di correlazione con trasformazione logaritmica: 0.2768891
- Indice di correlazione con trasformazione square root: 0.4139261
- Indice di correlazione con trasformazione quadratica: 0.2877589

Dagli indici di correlazione emerge come, in questo caso, sia meglio considerare la radice quadrata della variabile esplicativa Delicassen

Gli R^2 relativi ai tre modelli sono i seguenti:

- Semplice: 0.1713
- Con aggiunta della variabile Channel: 0.3377
- Con aggiunta dell'interazione con Channel: 0.3381

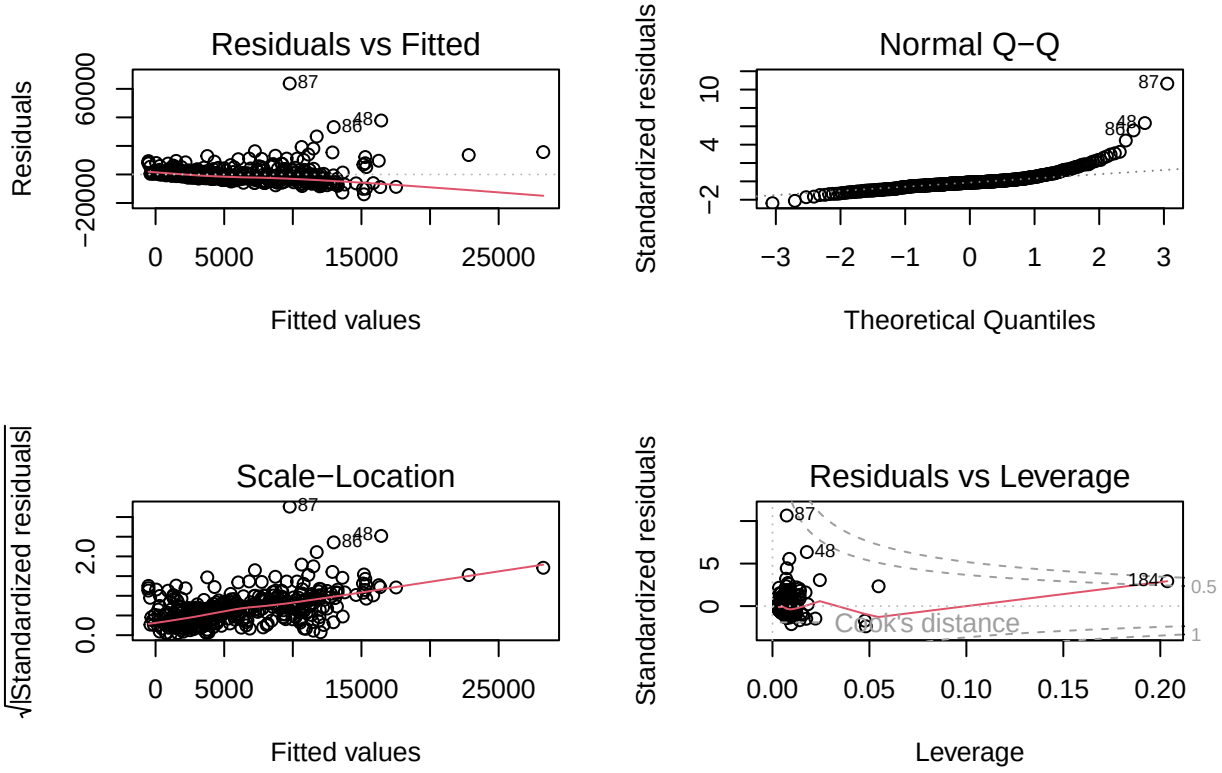
Output AIC ed ANOVA:

```
##       df      AIC
## fit14  3 9008.762
## fit24  4 8910.169
## fit34  5 8910.863
```

```
## Analysis of Variance Table
##
## Model 1: risp_var4 ~ esp_var4
## Model 2: risp_var4 ~ esp_var4 + Channel
## Model 3: risp_var4 ~ esp_var4 + Channel + esp_var4 * Channel
##   Res.Df        RSS Df  Sum of Sq       F Pr(>F)
## 1    438 1.9815e+10
## 2    437 1.5766e+10  1 4049644051 112.326 <2e-16 ***
## 3    436 1.5719e+10  1   46725517   1.296 0.2556
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

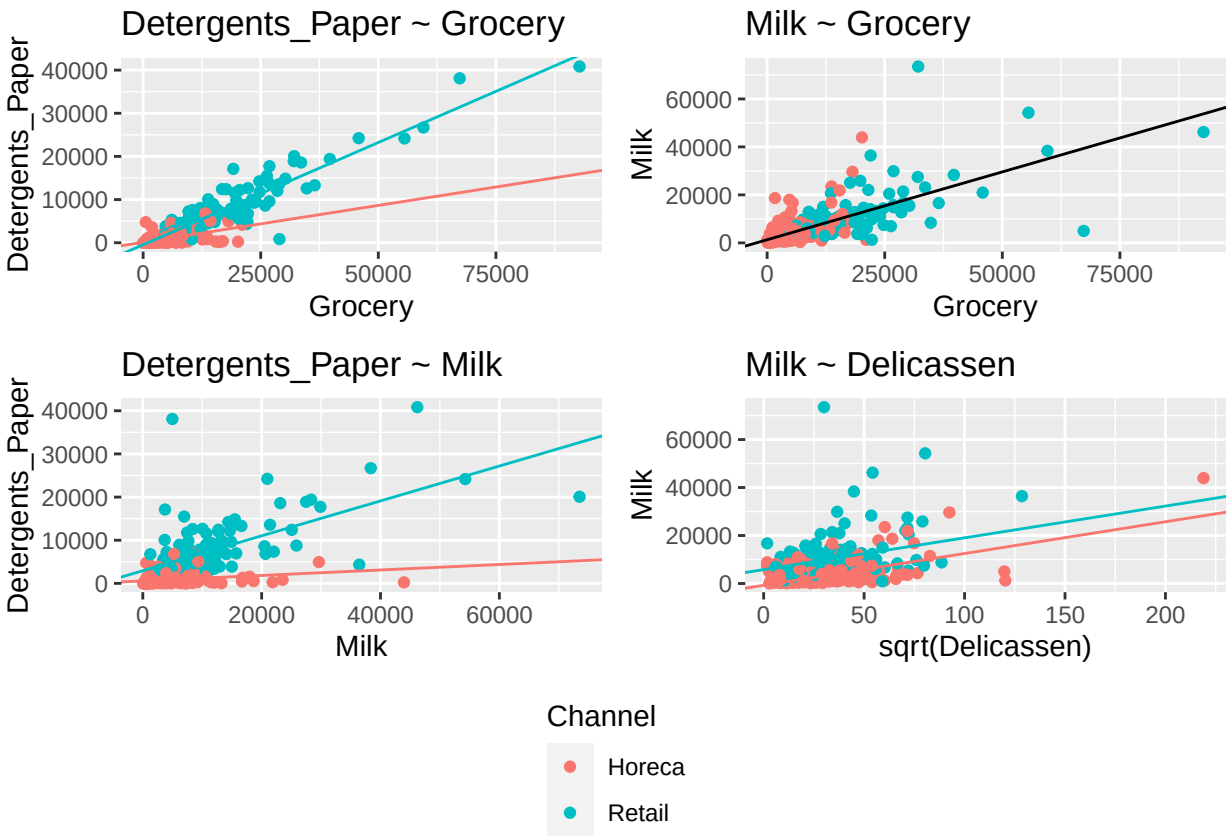
Sebbene i tre R^2 non siano alti, quindi i tre modelli non spiegano bene i dati, dopo aver valuatato anche i tre AIC e l'output del test ANOVA, emerge come il modello senza interazione ma con l'aggiunta di Channel sia il più adatto, in quanto, è vero che non è quello con R^2 maggiore, ma ha AIC minore e l'ANOVA mostra che l'aggiunta dell'interazione non è significativa (p-value = 0.2556).

Inoltre, per coerenza con quanto precedentemente detto, è preferibile utilizzare il modello più semplice possibile nel caso in cui le differenze a livello di R^2 e AIC non siano così rilevanti.



Anche in questo caso, comè successo con Milk ~ Grocery, nessuna ipotesi è rispettata, infatti osservando i grafici ed i risultati dell'oggetto gvlma, emerge come vi sia eteroschedasticità, non gaussianità, e non linearità nella relazione Milk condizionatamente a Delicassen.

Di seguito riportiamo i quattro scatterplot con le relative rette di regressione lineare:



Dagli scatterplot delle variabili quantitative, condizionatamente alla variabile categoriale Region, emerge come, al variare della regione, i punti sembrano mantenere lo stesso pattern, sebbene cambi la numerosità.

Per quanto riguarda gli scatterplot, condizionati alla variabile qualitativa Channel, si può notare come, cambiando il canale di vendita, i punti seguano pattern diversi, con ciò si può concludere che Channel è più influente sull'insieme dei dati rispetto a Region.

A seguito di queste considerazioni, nel momento in cui andremo a svillupare i modelli lineari, considereremo solo l'aggiunta della variabile categoriale Channel, ed, eventualmente, l'interazione della stessa con la variabile esplicativa.