Progetto finale corso di Foundations of Probability and Statistics

Susanna Maugeri 839365, Corinna Strada 839193

Pre-appello di dicembre 2021

Introduzione

Il dataset considerato, reperito su Kaggle, è disponibile al seguente **link**. Contiene dei dati su diverse composizioni di cemento e si compone di 9 variabili e 1030 osservazioni.

Le variabili di interesse sono:

- *Cement*: quantità di cemento, unità di misura kg/m^3 di miscela;
- *Blast Furnace Slag*: quantità di scorie in altoforno, unità di misura kg/m^3 di miscela;
- *Fly Ash*: quantità di cenere, unità di misura kg/m^3 di miscela;
- *Water*: quantità d'acqua, unità di misura kg/m^3 di miscela;
- *Superplasticizer*: quantità di superfluidificante, additivo in grado di ridurre la quantità d'acqua necessaria, unità di misura kg/m^3 di miscela;
- *Coarse Aggregate*: aggregato grossolano, unità di misura kg/m^3 di miscela;
- *Fine Aggregate*: aggregato fine, unità di misura kg/m^3 di miscela;
- *Age*: numero di giorni dopo i quali si testa la miscela, unità di misura: giorni 1~365;
- *Concrete Compressive Strength*: resistenza del calcestruzzo alla compressione, unità di misura MPa.

Per il task di regressione su cui ci concentreremo considereremo la variabile *Concrete Compressive Strength* come target e tutte le altre come regressori.

Importazione del dataset e controlli preliminari

```
##   Cement Blast.Furnace.Slag Fly.Ash Water Superplasticizer Coarse.Aggregate
## 1  540.0                0.0       0   162              2.5           1040.0
## 2  540.0                0.0       0   162              2.5           1055.0
## 3  332.5              142.5       0   228              0.0            932.0
## 4  332.5              142.5       0   228              0.0            932.0
## 5  198.6              132.4       0   192              0.0            978.4
## 6  266.0              114.0       0   228              0.0            932.0
##   Fine.Aggregate Age Strength
## 1          676.0  28    79.99
## 2          676.0  28    61.89
## 3          594.0 270    40.27
## 4          594.0 365    41.05
## 5          825.5 360    44.30
## 6          670.0  90    47.03
```

L'importazione è andata a buon fine.

Stampiamo un primo sommario delle variabili per verificare il numero di valori mancanti e se sono presenti variabili degeneri:

```
##             variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
## 1             Cement       0    0.00    0    0     0     0 numeric    278
## 2 Blast.Furnace.Slag     471   45.73    0    0     0     0 numeric    185
## 3            Fly.Ash     566   54.95    0    0     0     0 numeric    156
## 4              Water       0    0.00    0    0     0     0 numeric    195
## 5   Superplasticizer     379   36.80    0    0     0     0 numeric    111
## 6   Coarse.Aggregate       0    0.00    0    0     0     0 numeric    284
## 7     Fine.Aggregate       0    0.00    0    0     0     0 numeric    302
## 8                Age       0    0.00    0    0     0     0 numeric     14
## 9           Strength       0    0.00    0    0     0     0 numeric    845
```

Nessuna variabile presenta valori mancanti e tutte sono di tipo numerico.

Calcoliamo le statistiche descrittive principali per le variabili del dataset e la matrice di correlazione.

```
##      Cement      Blast.Furnace.Slag    Fly.Ash           Water      
##  Min.   :102.0   Min.   :  0.0      Min.   :  0.00   Min.   :121.8  
##  1st Qu.:192.4   1st Qu.:  0.0      1st Qu.:  0.00   1st Qu.:164.9  
##  Median :272.9   Median : 22.0      Median :  0.00   Median :185.0  
##  Mean   :281.2   Mean   : 73.9      Mean   : 54.19   Mean   :181.6  
##  3rd Qu.:350.0   3rd Qu.:142.9      3rd Qu.:118.30   3rd Qu.:192.0  
##  Max.   :540.0   Max.   :359.4      Max.   :200.10   Max.   :247.0  
##  Superplasticizer Coarse.Aggregate Fine.Aggregate       Age        
##  Min.   : 0.000   Min.   : 801.0   Min.   :594.0   Min.   :  1.00  
##  1st Qu.: 0.000   1st Qu.: 932.0   1st Qu.:731.0   1st Qu.:  7.00  
##  Median : 6.400   Median : 968.0   Median :779.5   Median : 28.00  
##  Mean   : 6.205   Mean   : 972.9   Mean   :773.6   Mean   : 45.66  
##  3rd Qu.:10.200   3rd Qu.:1029.4   3rd Qu.:824.0   3rd Qu.: 56.00  
##  Max.   :32.200   Max.   :1145.0   Max.   :992.6   Max.   :365.00  
##     Strength    
##  Min.   : 2.33  
##  1st Qu.:23.71  
##  Median :34.45  
##  Mean   :35.82  
##  3rd Qu.:46.13  
##  Max.   :82.60
```

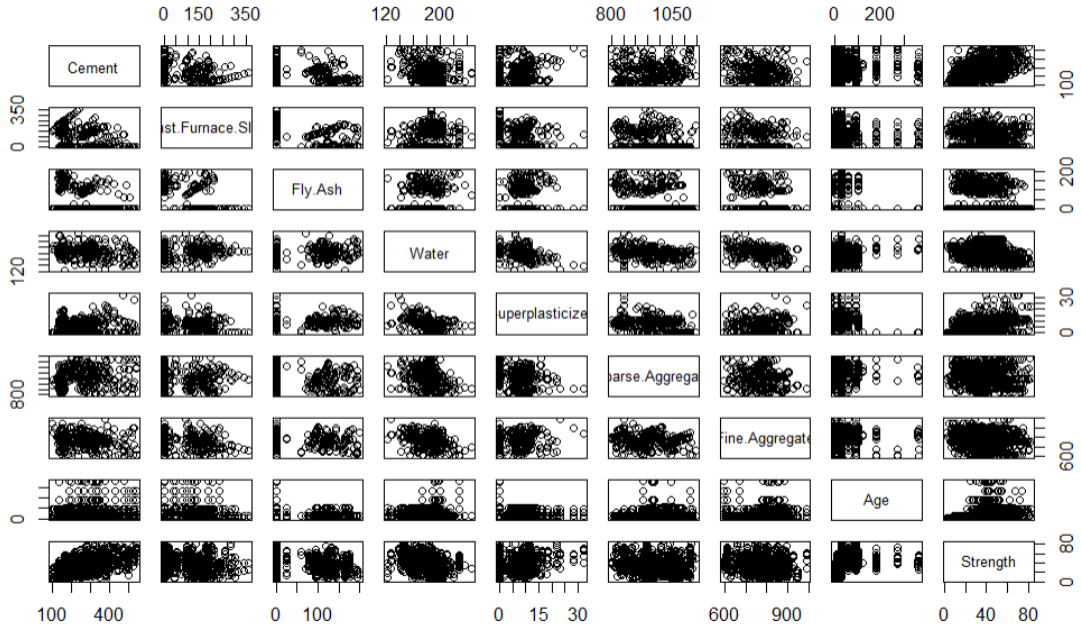
```
##                    vars    n   mean     sd median trimmed    mad    min    max
## Cement                1 1030 281.17 104.51 272.90  273.47 117.72 102.00  540.0
## Blast.Furnace.Slag    2 1030  73.90  86.28  22.00   62.43  32.62   0.00  359.4
## Fly.Ash               3 1030  54.19  64.00   0.00   46.86   0.00   0.00  200.1
## Water                 4 1030 181.57  21.35 185.00  181.19  19.27 121.80  247.0
## Superplasticizer      5 1030   6.20   5.97   6.40    5.56   7.86   0.00   32.2
## Coarse.Aggregate      6 1030 972.92  77.75 968.00  973.49  68.64 801.00 1145.0
## Fine.Aggregate        7 1030 773.58  80.18 779.50  776.41  67.46 594.00  992.6
## Age                   8 1030  45.66  63.17  28.00   32.53  31.13   1.00  365.0
## Strength              9 1030  35.82  16.71  34.44   34.96  16.20   2.33   82.6
```

```
##                      range  skew kurtosis   se
## Cement              438.00  0.51    -0.53 3.26
## Blast.Furnace.Slag  359.40  0.80    -0.52 2.69
## Fly.Ash             200.10  0.54    -1.33 1.99
## Water               125.20  0.07     0.11 0.67
## Superplasticizer     32.20  0.90     1.39 0.19
## Coarse.Aggregate    344.00 -0.04    -0.61 2.42
## Fine.Aggregate      398.60 -0.25    -0.11 2.50
## Age                 364.00  3.26    12.07 1.97
## Strength             80.27  0.42    -0.32 0.52
```

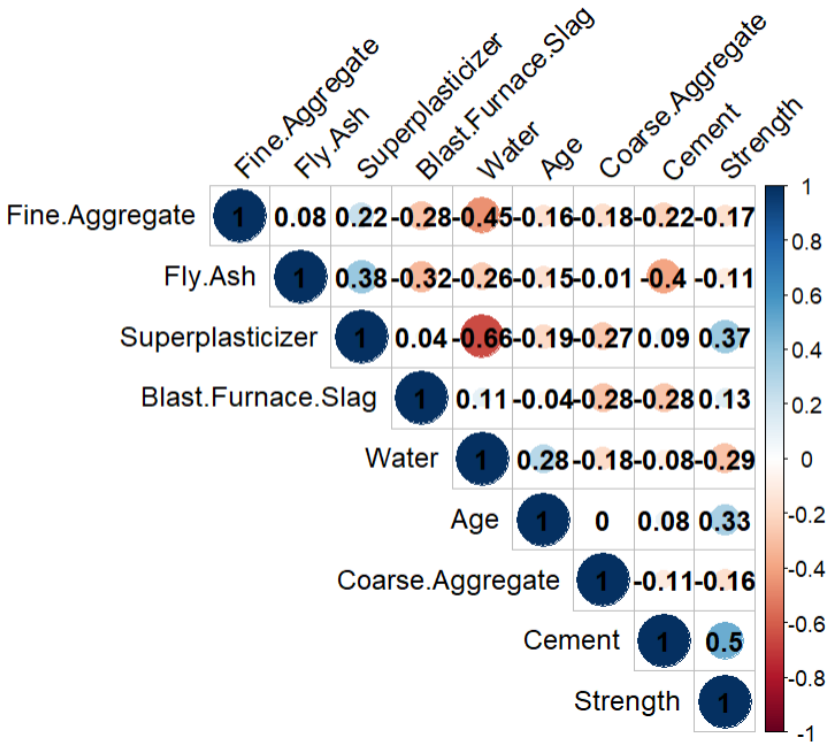
Per ogni variabile sono riportate le statistiche descrittive più importanti. Notiamo in particolare che tutte le variabili, ad accezione di *Age*, non presentano asimmetrie particolarmente importanti. La simmetria massima si registra relativamente a *Coarse.Aggregate* (-0.04) e quell'attributo, insieme a *Fine.Aggregate*, è l'unico a presentare un valore dell'indice negativo. La massima asimmetria, di tipo positivo, si registra invece per la variabile *Age*, la quale infatti presenta valori di media aritmetica e mediana molto distanti tra di loro.

Per quanto riguarda la curtosi, anche in questo caso il valore massimo è registrato da *Age* ed è pari a 12.07.Ciò significa che essa ha code definite come "pesanti" e la sua distribuzione è leptocurtica. Tutti gli altri valori non si distaccano troppo dallo 0, quindi le altre variabili sembrano presentare una distribuzione che ha una forma non troppo differente da quella normale per asimmetria.

Plottiamo la matrice dei diagrammi di dispersione per avere una idea sulla distribuzione delle variabili e successivamente la matrice di correlazione.



A causa del numero elevato di variabili e di osservazioni, è difficile osservare la distribuzione dei punti nei grafici, per questo produciamo la matrice di correlazione.

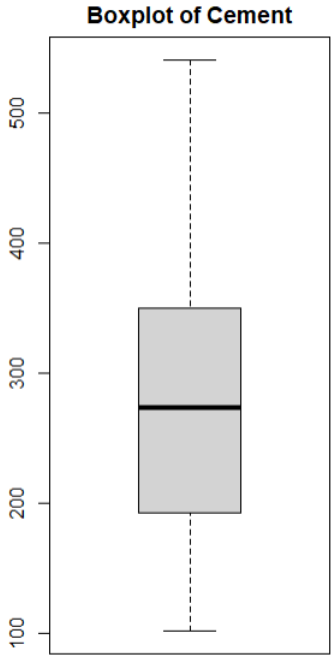


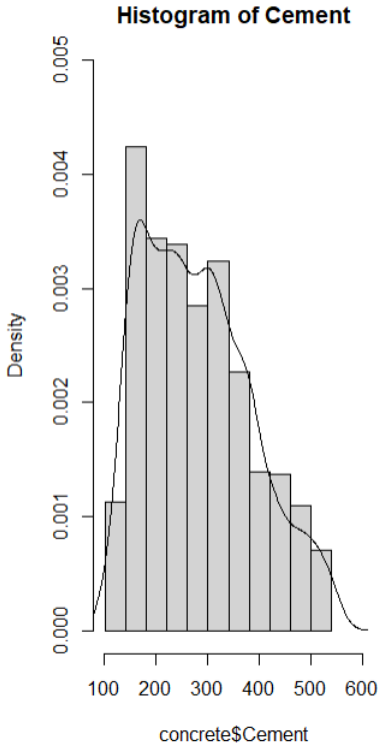
Dalla matrice di correlazione si può vedere che le variabili *Water* e *Superplasticizer* presentano una correlazione negativa e piuttosto elevata, pari a -0.66, così come *Fine.Aggregate* e *Water* pari a 0.45. Le variabili più correlate con il target *Strength* sono *Cement* (0.5), *Superplasticizer* e *Age* (0.33): si può dunque ipotizzare, ad una prima superficiale analisi, che il calcestruzzo tenda ad essere maggiormente resistente alla compressione se viene utilizzata una maggior quantità di cemento e una minore di superfluidificante e in generale al passare del tempo.

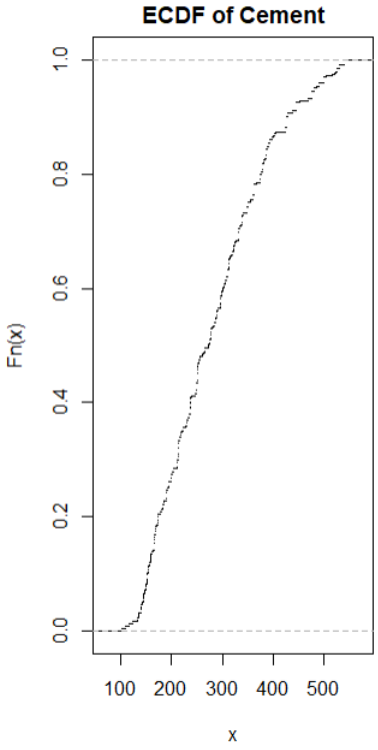
Analisi delle variabili

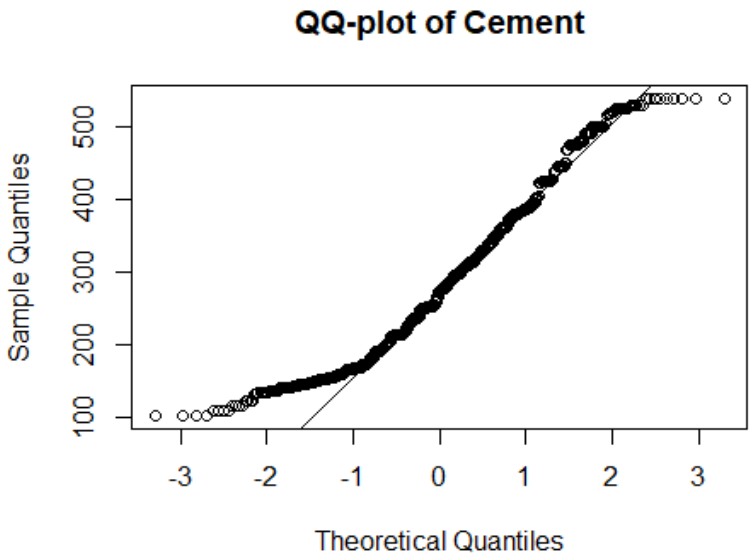
Per ciascuna variabile vengono prodotti il boxplot, l'istogramma, sul quale si vede anche la funzione di densità dell'attributo, il plot della funzione di ripartizione empirica e il QQ-plot e infine viene effettuato il test di Shapiro-Wilk per testare l'ipotesi di normalità delle loro distribuzioni.

Variabile *Cement*







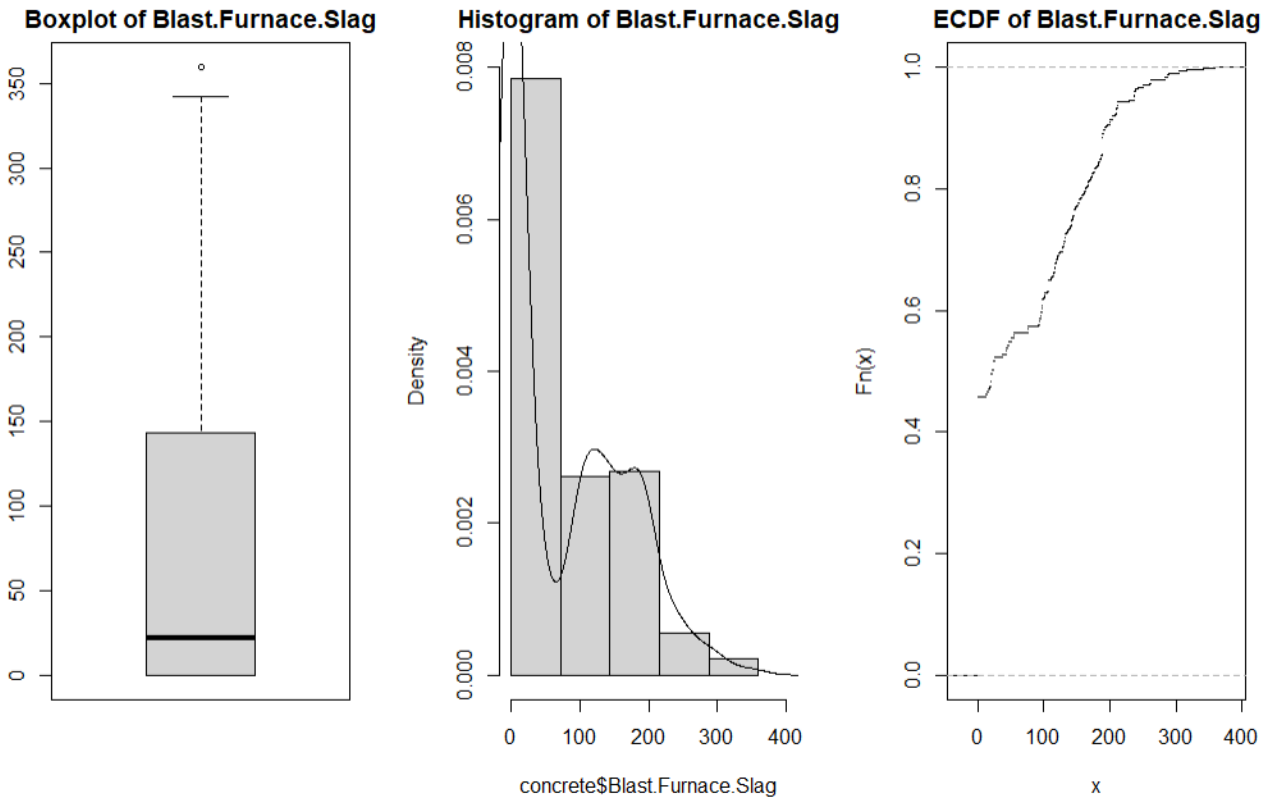


La distribuzione di *Cement* risulta leggermente asimmetrica a destra, come già in precedenza osservato. Infatti, nel summary del dataframe si osserva che la mediana è maggiore della media di pochi punti e che esse si trovano non centrate rispetto al primo e al terzo quartile della distribuzione. Dal boxplot non risultano esserci valori inusuali, all'infuori cioè dei baffi che rappresentano il Range Interquartile.

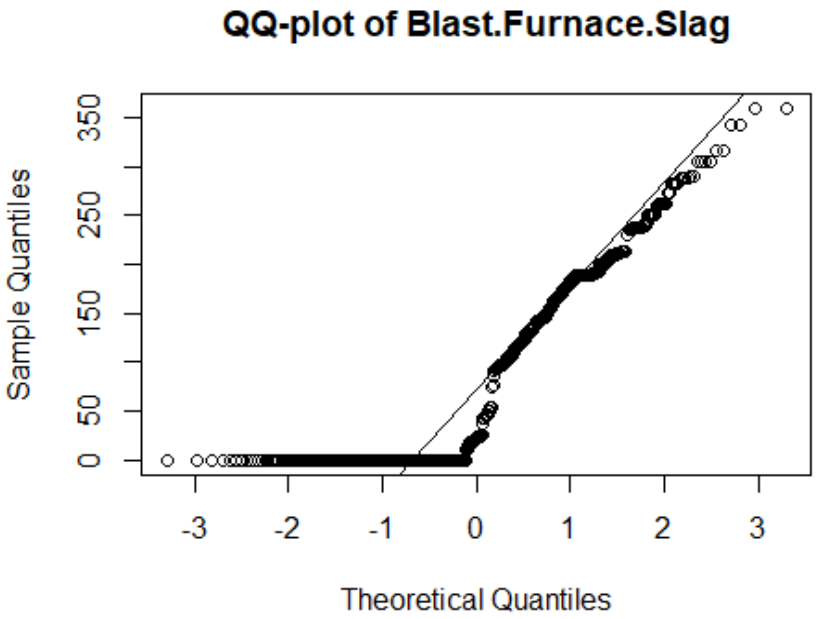
```
##
##  Shapiro-Wilk normality test
##
## data:  concrete$Cement
## W = 0.95896, p-value < 2.2e-16
```

Il QQ-plot ed il test di Shapiro-Wilk confermano l'ipotesi di non normalità: le code della distribuzione, specialmente quella di sinistra, risultano molto spostate rispetto ai quantili teorici.

Variabile *Blast.Furnace.Slag*



Nel Boxplot si osserva un pallino fuori dal baffo di destra, ad un'analisi più dettagliata abbiamo scoperto che si tratta in realtà di due osservazioni, la 554 e la 560, che presentano il medesimo valore 359.4 per questa variabile.

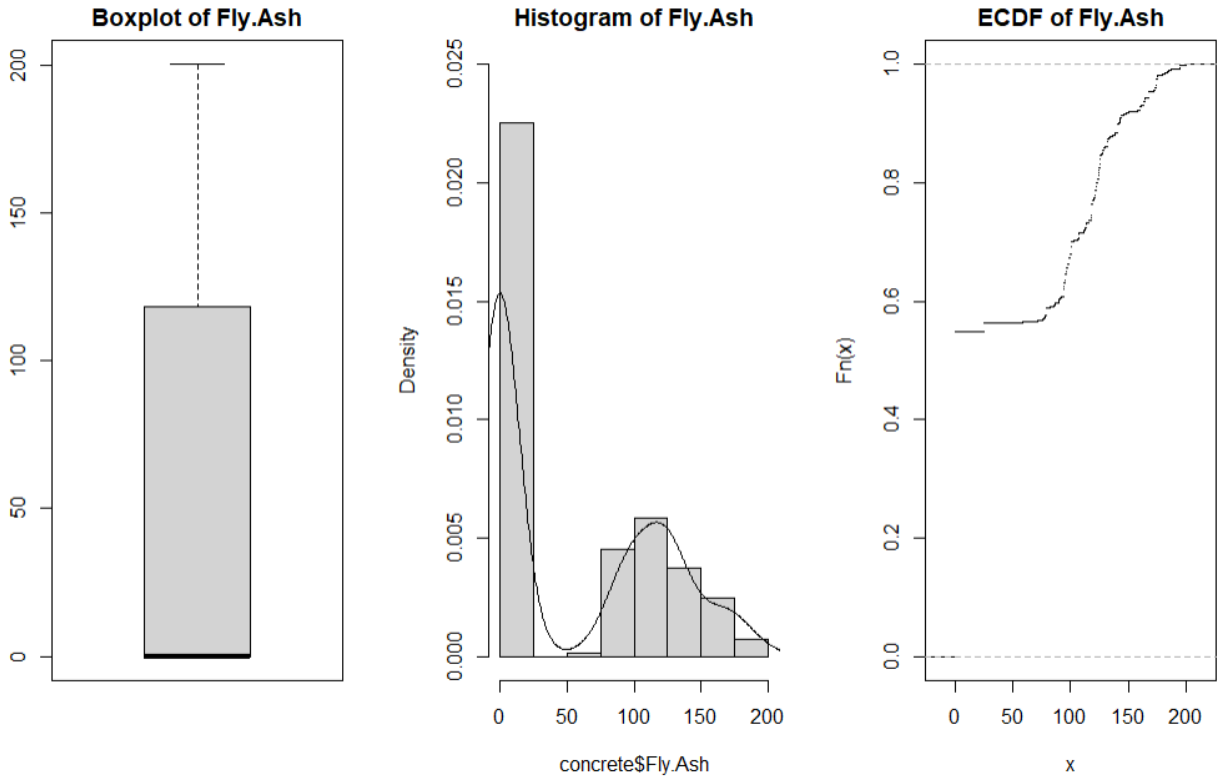


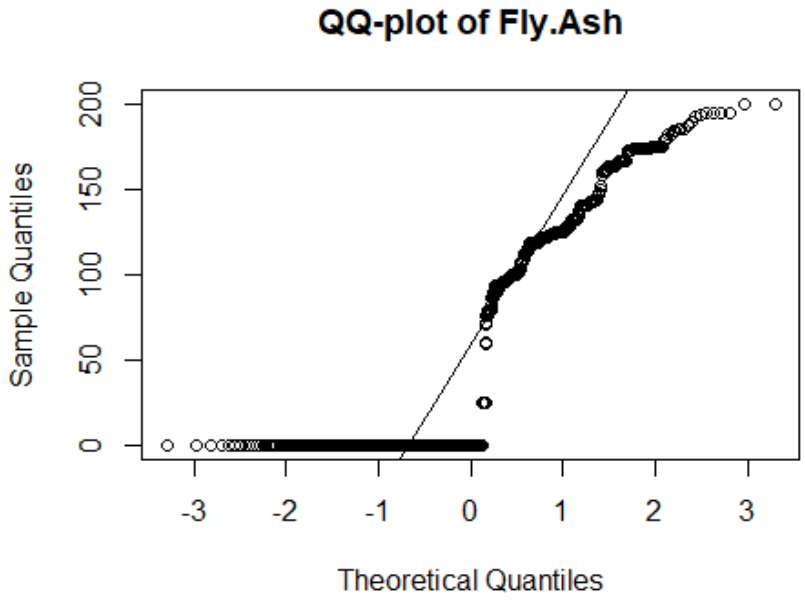
La distribuzione di *Blast Furnace Slag* risulta abbastanza asimmetrica a destra. Anche in questo caso ciò era intuibile dai quantili calcolati precedentemente: il primo quartile ha valore 0, perciò almeno il 25% dei dati presenta valore nullo. Tuttavia dal plot della funzione di ripartizione empirica si nota che più del 40% dei dati in realtà presenta valore nullo. Inoltre dal boxplot si vede che la variabile presenta un'osservazione outlier nell coda destra della distribuzione.

```
##
##  Shapiro-Wilk normality test
##
## data:  concrete$Blast.Furnace.Slag
## W = 0.81241, p-value < 2.2e-16
```

Il QQ-plot e il test di Shapiro-Wilk confermano l'ipotesi di non normalità della variabile.

Variabile *Fly.Ash*

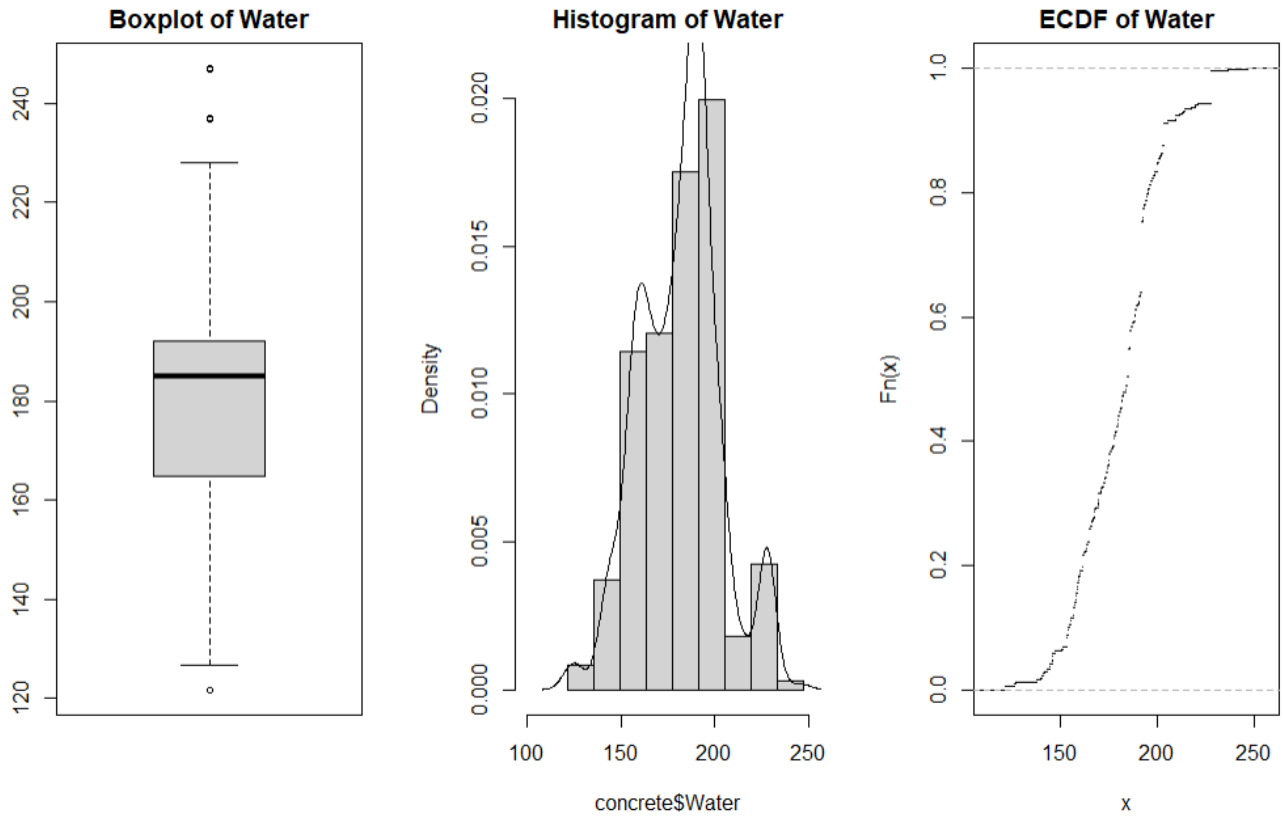




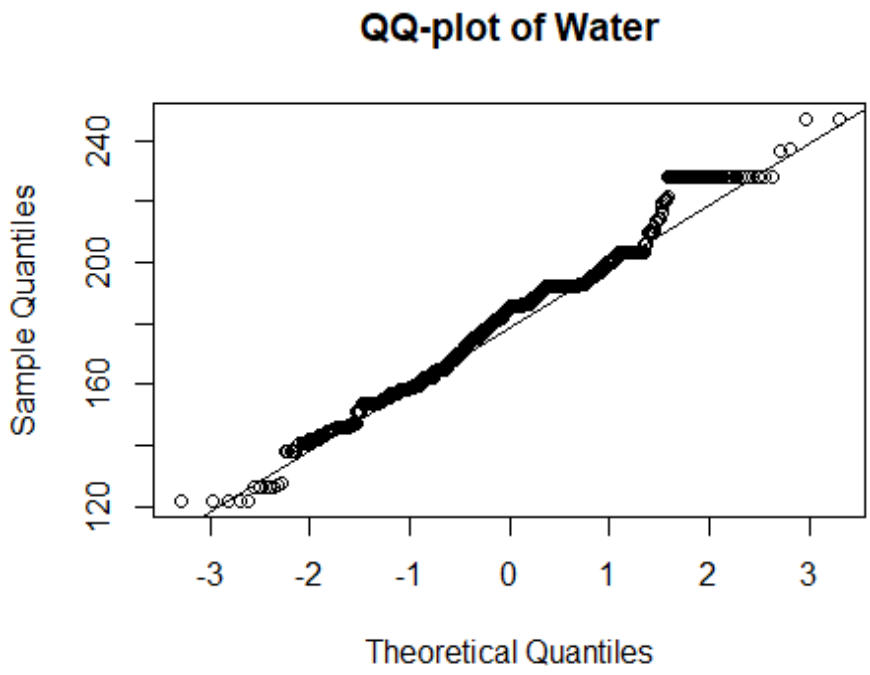

```
##
##  Shapiro-Wilk normality test
##
## data:  concrete$Fly.Ash
## W = 0.762, p-value < 2.2e-16
```

La distribuzione di *Fly Ash* risulta molto asimmetrica a destra. Dai quantili calcolati infatti si legge che più della metà dei dati presenta valore nullo; in effetti dalla funzione di ripartizione empirica si vede che quasi il 60% dei dati ha valore nullo. Dall'istogramma si vede che la variabile ha una distribuzione molto particolare: una decisa prevalenza di valori nulli e una distribuzione dei valori non nulli a sua volta asimmetrica leggermente a destra. Il Boxplot mostra che non vi sono valori che si posizionano all'infuori del Range Interquartile. Anche questa volta il QQ-plot e il test di Shapiro-Wilk confermano che la variabile non ha un andamento normale.

Variabile *Water*



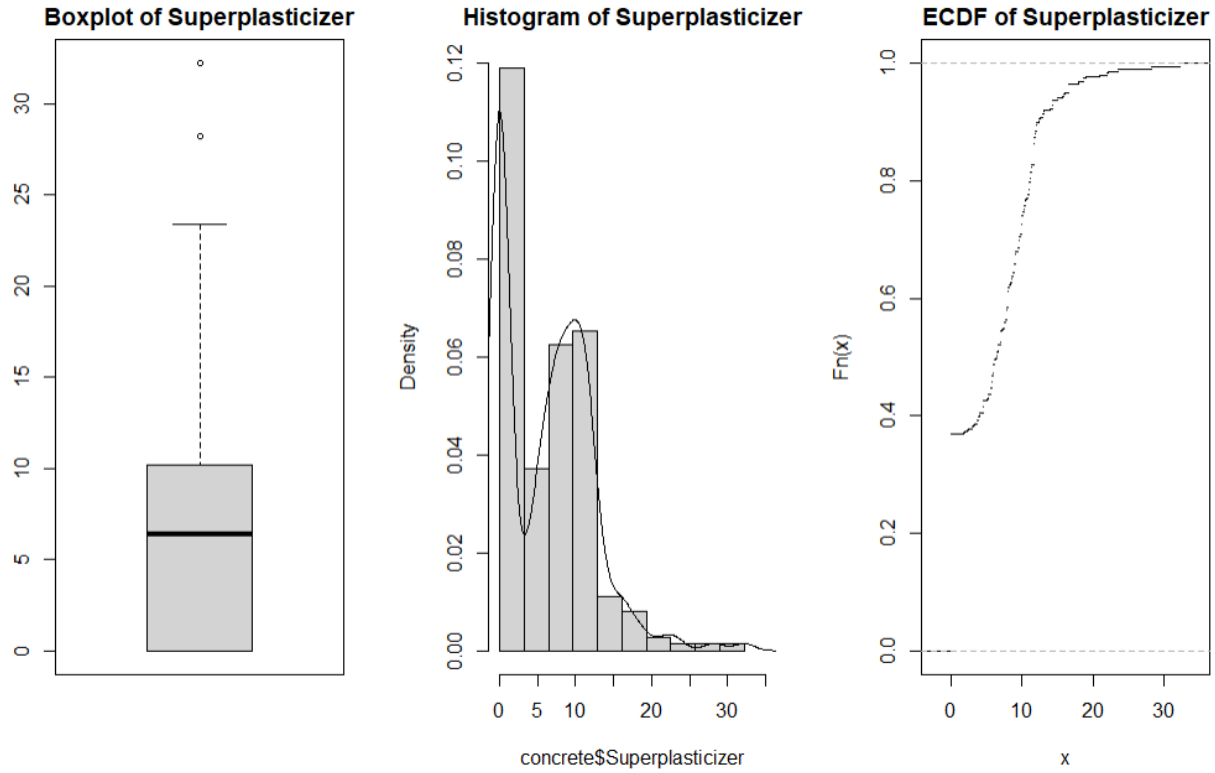
Il Boxplot mostra che vi sono dei valori all'infuori del Range Interquartile sia nella coda di destra che in quella di sinistra. Ad un'analisi più approfondita abbiamo scoperto che si tratta delle osservazioni dalla 225 alla 229 per la coda di sinistra, che presentano tutte un valore di *Water* pari a 121.8, e delle 863, 874, 937 e 1020 per la coda di destra. Dall'istogramma la variabile sembra essere leggermente asimmetrica a sinistra.



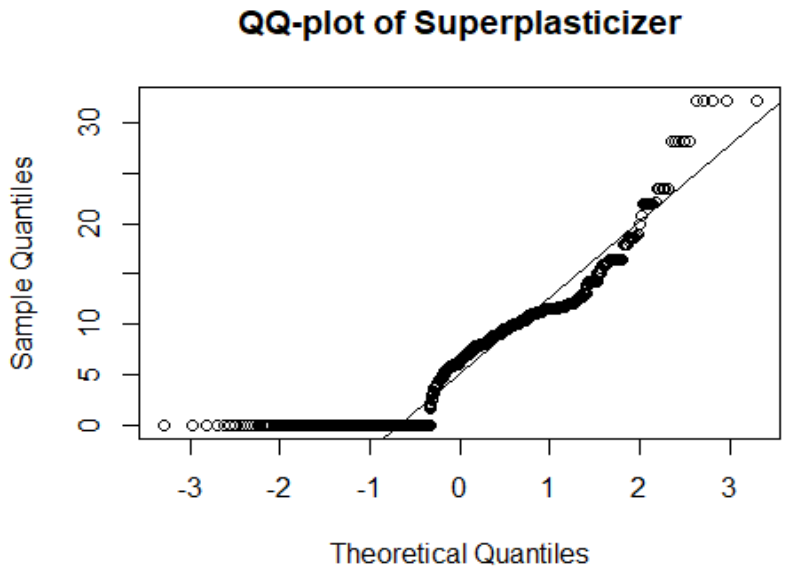

```
##
##  Shapiro-Wilk normality test
##
## data:  concrete$Water
## W = 0.98039, p-value = 1.463e-10
```

Il QQ-plot mostra che la variabile ha un andamento quasi normale, ad eccezione della coda di destra in cui i quantili osservati e teorici differiscono notevolmente.

Variabile *Superplasticizer*



Anche la variabile *Superplasticizer* risulta, in distribuzione, asimmetrica a destra. Il boxplot mostra, oltre all'asimmetria, anche che vi sono nella coda di destra dei valori outliers: si osservano due palline fuori dal baffo; ad un'analisi più approfondita risulta che si tratta di 10 osservazioni, delle quali 5 presentano un valore di *Superplasticizer* pari a 28.2 e 5 un valore di 32.2. Il plot della funzione di ripartizione empirica, infine, mostra che quasi il 40% dei dati ha valore nullo.

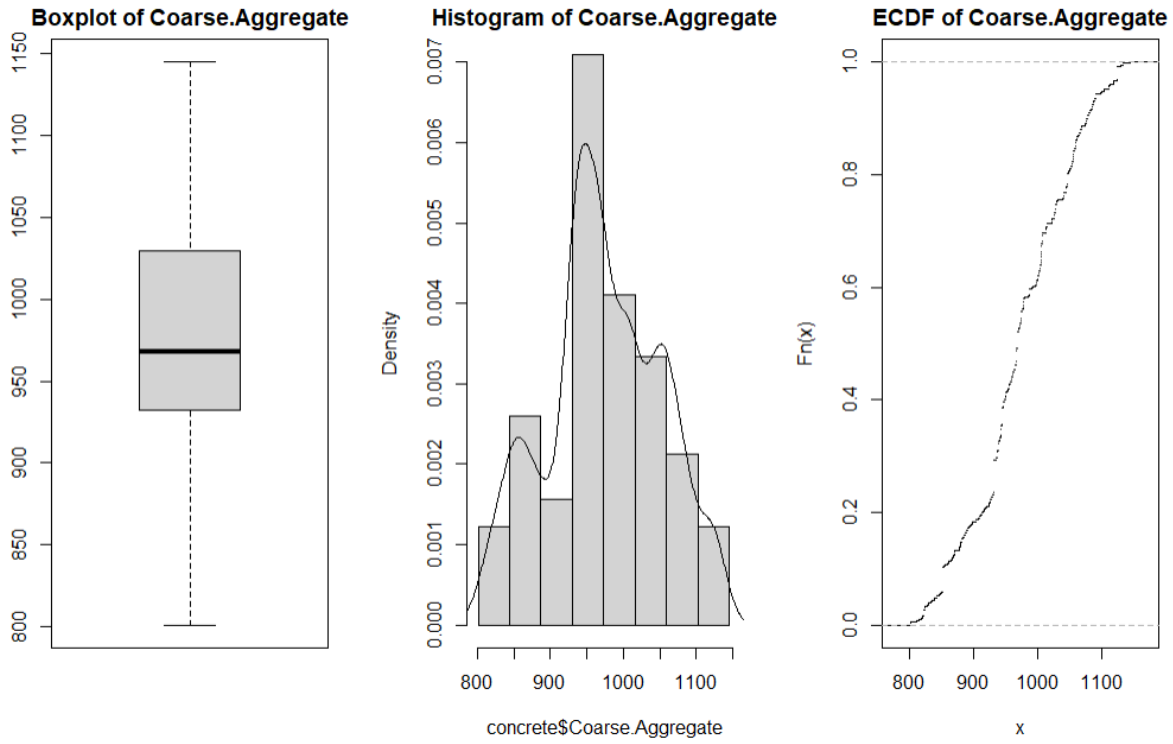


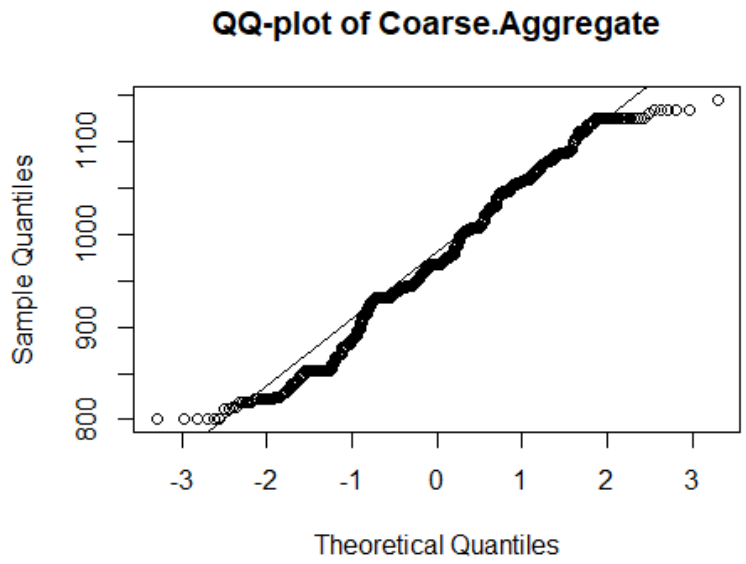
Anche per questa variabile il QQ-plot mostra un andamento totalmente non normale.

```
##
##  Shapiro-Wilk normality test
##
## data:  concrete$Superplasticizer
## W = 0.86603, p-value < 2.2e-16
```

Il test di Shapiro-Wilk conferma l'ipotesi di non normalità fatta osservando i grafici precedenti.

Variabile *Coarse.Aggregate*



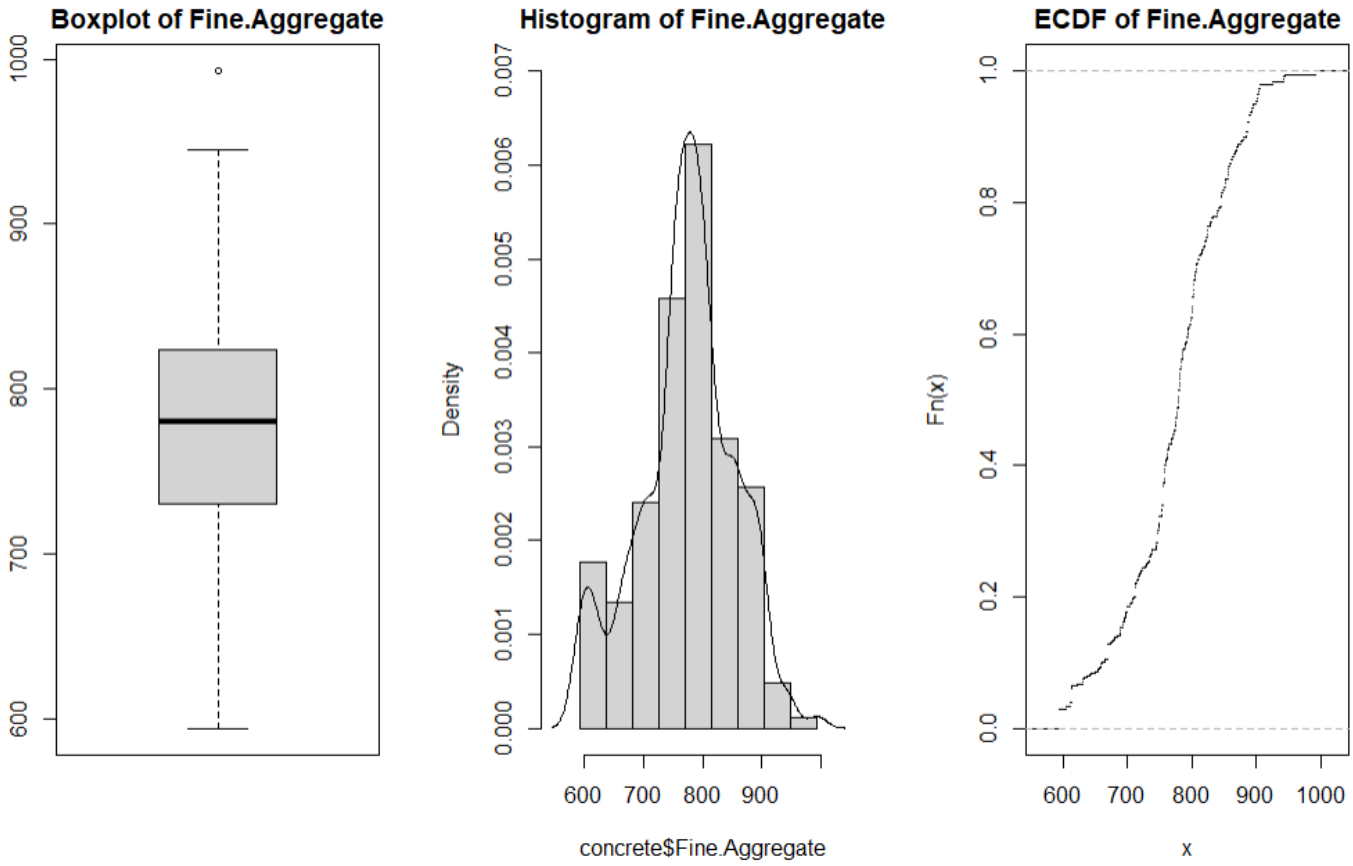


Dai grafici bBxplot e Istogramma la variabile sembra essere abbastanza simmetrica, tuttavia il QQ-plot confuta questa idea mostrando un andamento che non coincide con quello normale. Il Boxplot mostra inoltre che non ci sono valori che si posizionano fuori dal Range Interquartile.

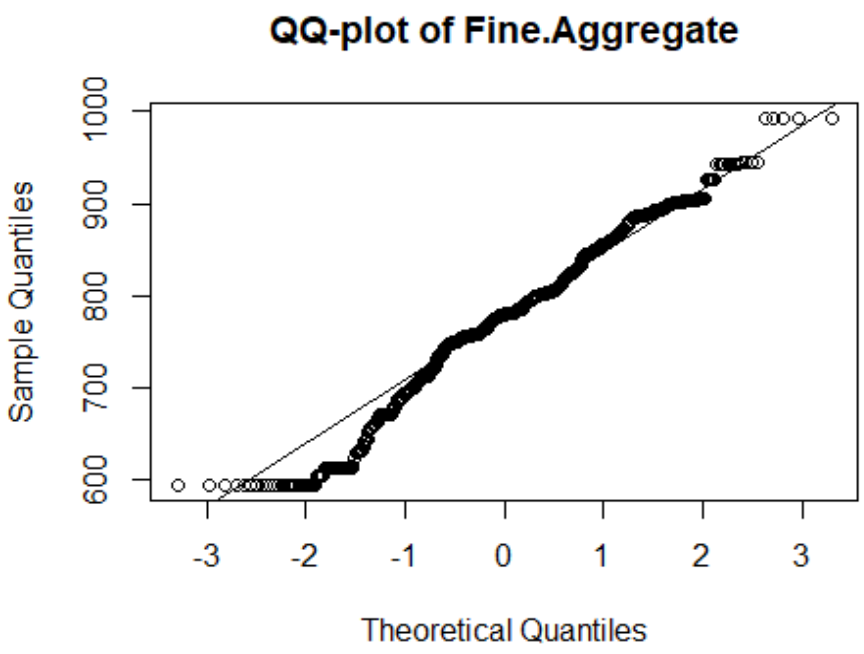
```
##
##  Shapiro-Wilk normality test
##
## data:  concrete$Coarse.Aggregate
## W = 0.98245, p-value = 8.347e-10
```

Anche in questo caso il test di Shapiro-Wilk conferma la non normalità della distribuzione.

Variabile *Fine.Aggregate*



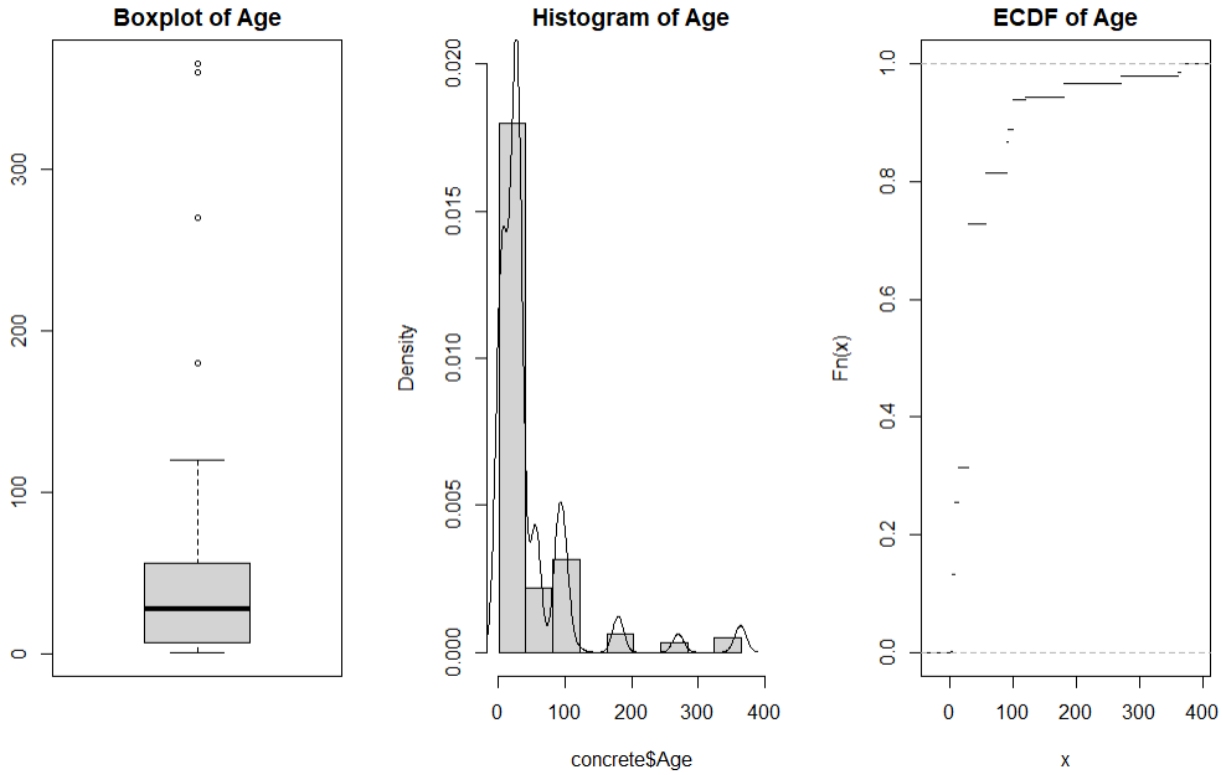
Il Boxplot mostra un pallino fuori dal baffo di destra, in realtà si tratta di 5 osservazioni che presentano il medesimo valore 992.6 per la variabile considerata.



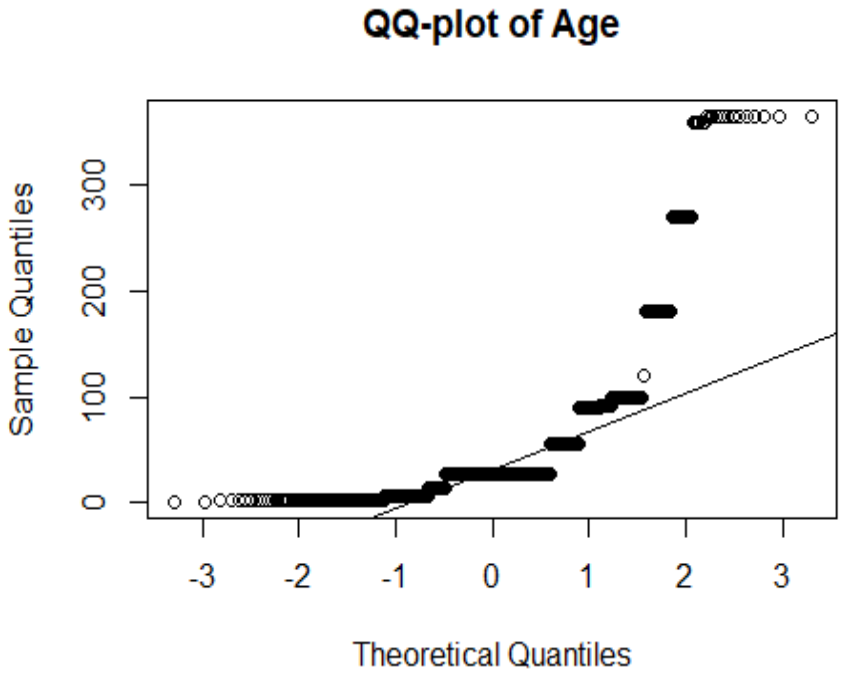

```
##
##  Shapiro-Wilk normality test
##
## data:  concrete$Fine.Aggregate
## W = 0.98067, p-value = 1.842e-10
```

Dall'istogramma la variabile sembra essere leggermente asimmetrica a sinistra, mentre dal QQ-plot si vede che la coda di sinistra è molto spostata rispetto ai quantili teorici della normale. Infine, con il test di Shapiro-Wilk viene confermato che la variabile non ha un andamento normale.

Variabile *Age*



I grafici per la variabile *Age* mostrano è essa è molto asimmetrica a destra e che la quasi totalità delle osservazioni presenta valori minori di 120, che è anche il valore estremo del baffo di destra del Boxplot. I valori fuori dal baffo rappresentano 59 osservazioni che riportano per la variabile i valori 180, 270, 360 e 365.

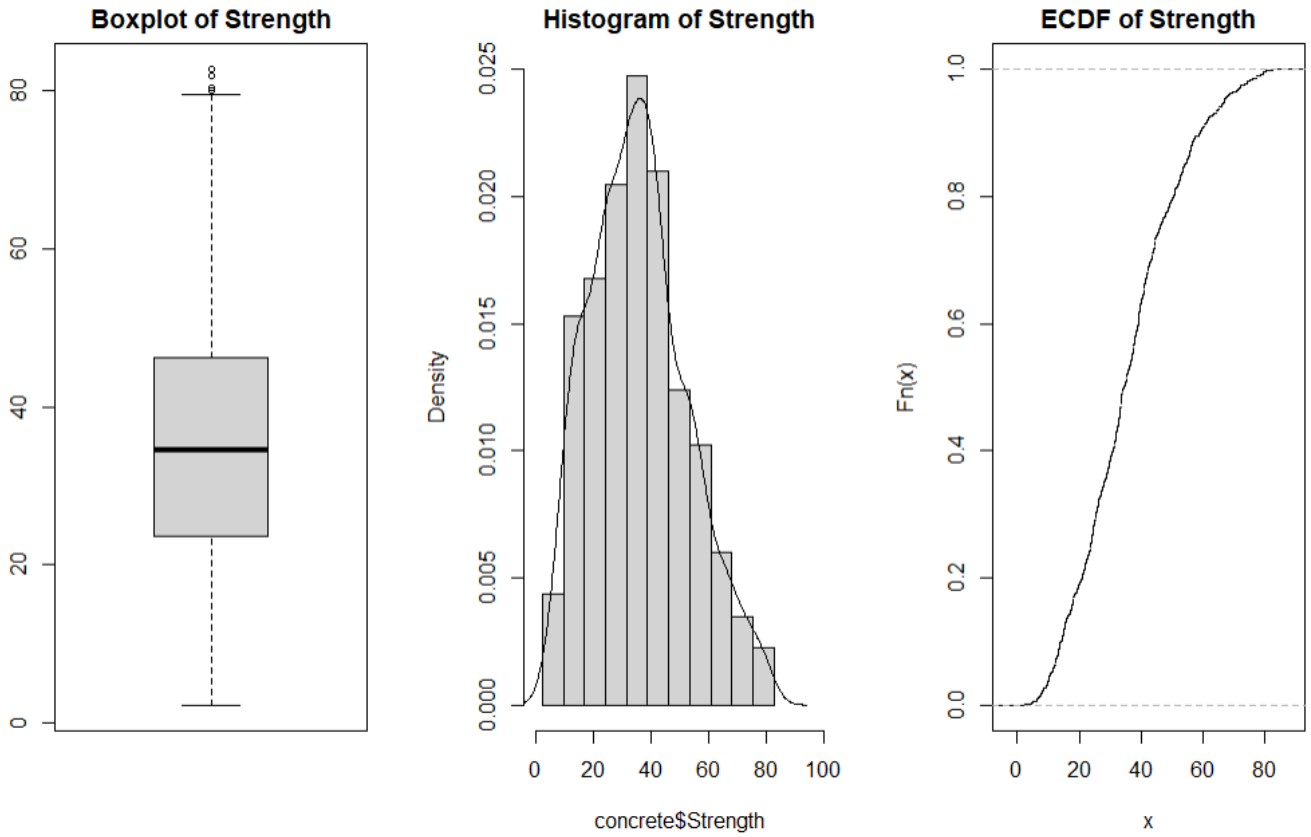


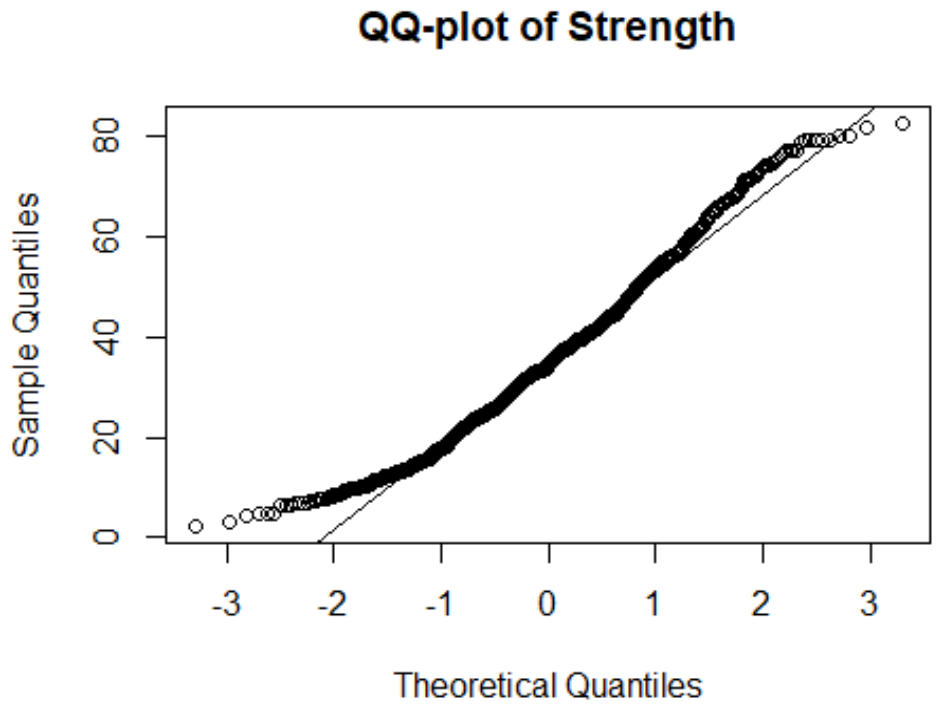
Il QQ-plot conferma sia l'ipotesi iniziale che la variabile non presenti assolutamente un andamento normale. Si notano dei segmenti orizzontali a causa del fatto che la variabile sia discreta e non continua.

```
##
##  Shapiro-Wilk normality test
##
## data:  concrete$Age
## W = 0.59071, p-value < 2.2e-16
```

Il bassissimo valore del p-value conferma l'ipotesi di non normalità della variabile.

Variabile *Strength*





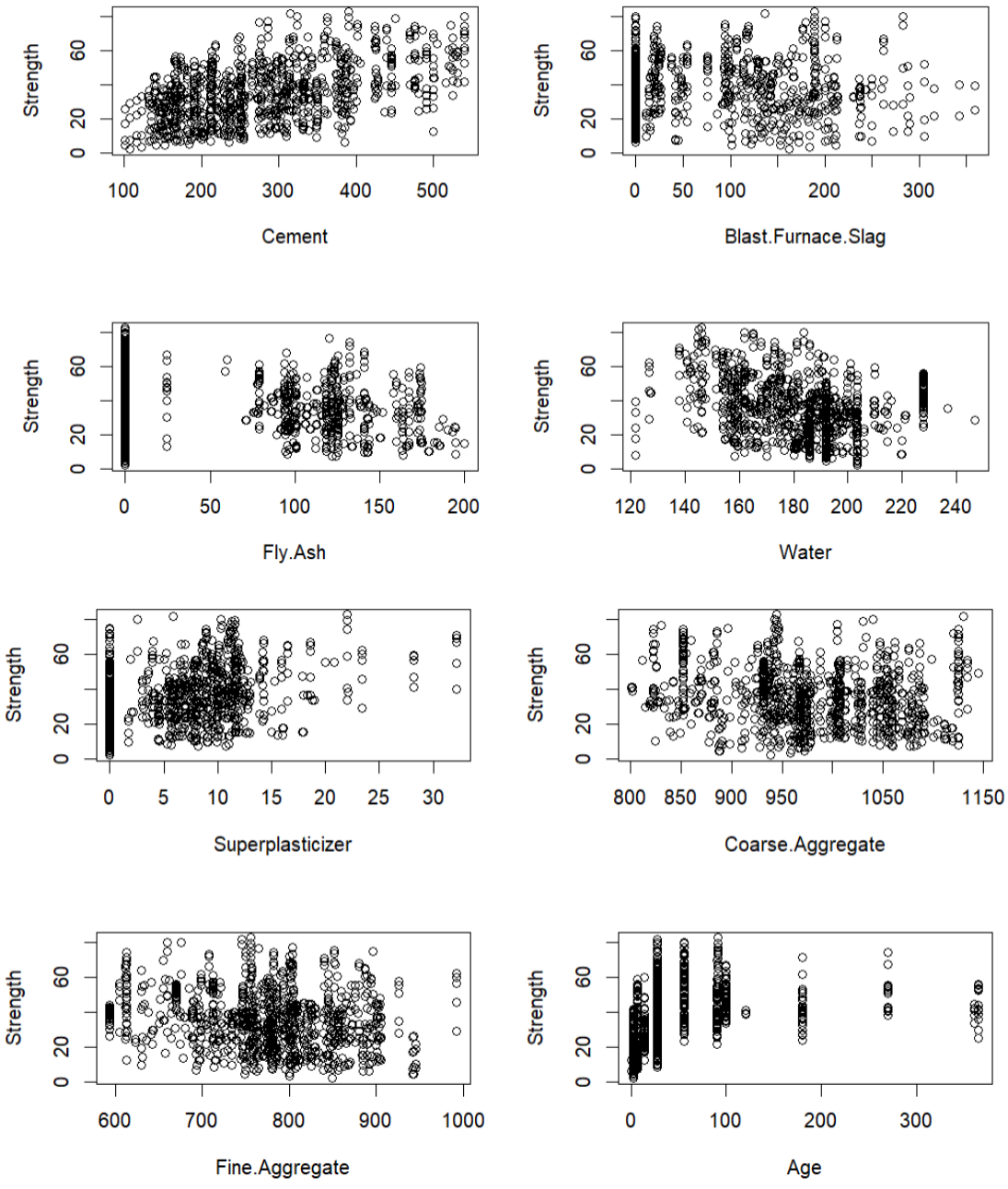
Infine, anche la variabile *Strength* risulta asimmetrica a destra, con 4 outliers nella coda di destra visibili nel Boxplot, che sono le osservazioni 1, 154, 182 e 382. Il QQ-plot mostra tuttavia un andamento abbastanza normale, ad eccezione delle code.

```
##
##  Shapiro-Wilk normality test
##
## data:  concrete$Strength
## W = 0.97979, p-value = 9.01e-11
```

Tuttavia, poiché il test di Shapiro-Wilk indica che l'ipotesi di normalità sia da rifiutare, vuol dire che le code hanno un peso notevole nella distribuzione, che risulta infatti non normale.

Diagrammi di dispersione dei singoli predittori rispetto alla variabile risposta

Plottiamo ora i diagrammi di dispersione dei singoli predittori rispetto alla variabile risposta *Strength*.



Sembra non esserci alcun tipo di legame lineare tra le singole variabili e la risposta, eccezion fatta per la covariata *Cement*. In quel caso, si registra una relazione lineare di tipo positivo: all'aumentare del valore di *Cement*, anche *Strength* sembra aumentare e viceversa, come notato precedentemente nella matrice di correlazione, dove la correlazione tra le due variabili si mostrava essere pari a 0.5. Si nota poi una leggera correlazione negativa tra la variabile *Water* e la variabile risposta; anche questa relazione era stata precedentemente notata nella matrice di correlazione, che presentava per queste due variabili un coefficiente pari a -0.29.

I grafici sono indicativi solo della relazione fra la variabile risposta e ciascuna delle variabili esplicative, ma non danno informazione globale sulla dipendenza della risposta da tutte le variabili esplicative considerate simultaneamente. È importante considerare anche la possibile dipendenza tra coppie di variabili esplicative, ma ciò viene tenuto in considerazione nella matrice di correlazione.

Costruzione del modello di regressione lineare multipla

Per la costruzione del miglior modello di regressione lineare, abbiamo suddiviso il dataset in due parti, il training set e il test set. Il training set contiene l'80% delle osservazioni del dataset iniziale e viene utilizzato per il fitting dei modelli. Il test set, invece, contiene la rimanente parte di osservazioni e su di esso sono stati applicati i modelli generati al fine di fare diverse considerazioni sui valori che essi predicono.

Prima di effettuare la partizione scegliamo un seed pari a 1234 per rendere possibile l'esatta replicazione del risultato.

Con questa suddivisione otteniamo un training set con 826 osservazioni ed un test set che ne comprende 204.

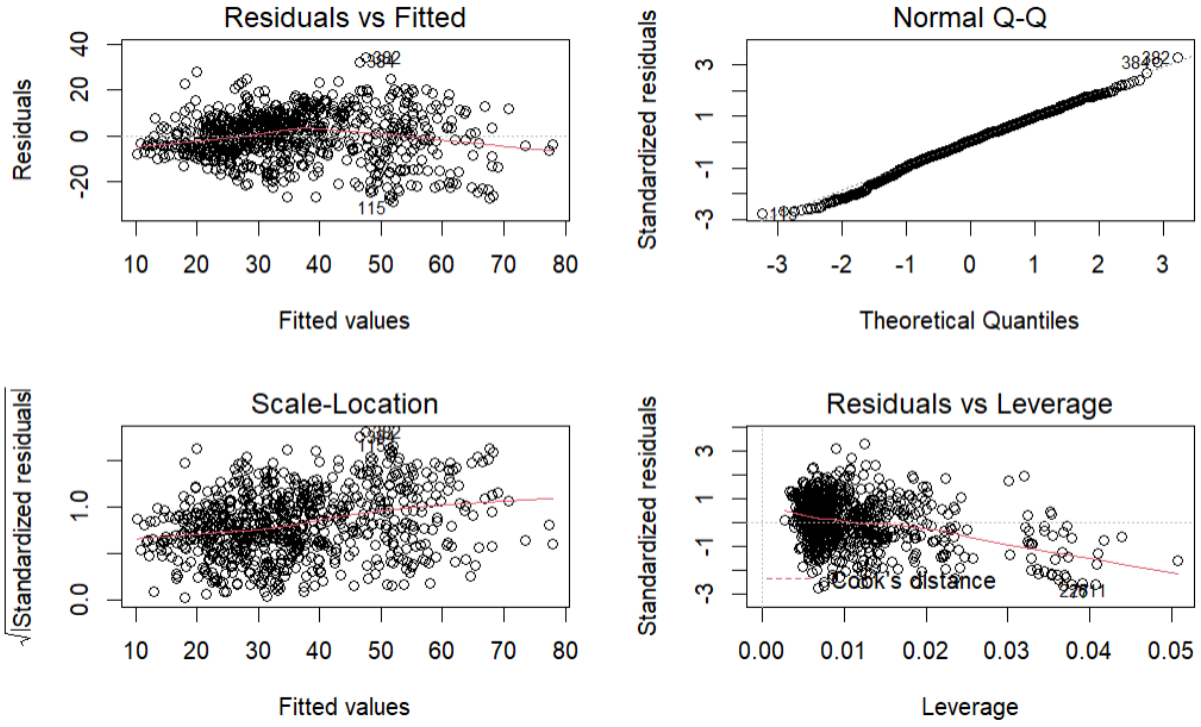
Il primo modello stimato, chiamato mod1, è quello completo, che utilizza cioè tutte le covariate. È stato costruito tramite il metodo chiamato lm ed è stato poi stampato il summary ad esso relativo.

```
##
## Call:
## lm(formula = Strength ~ ., data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.992  -6.343   0.564   7.067  34.168
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -26.701694  29.360443  -0.909 0.363384
## Cement               0.120688   0.009322  12.947  < 2e-16 ***
## Blast.Furnace.Slag   0.104093   0.011235   9.265  < 2e-16 ***
## Fly.Ash              0.089939   0.013962   6.442 2.01e-10 ***
## Water               -0.150356   0.044196  -3.402 0.000701 ***
## Superplasticizer     0.324697   0.103877   3.126 0.001836 **
## Coarse.Aggregate     0.019595   0.010369   1.890 0.059144 .
## Fine.Aggregate       0.022364   0.011943   1.872 0.061496 .
## Age                  0.110319   0.006020  18.325  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.49 on 817 degrees of freedom
## Multiple R-squared:  0.6086, Adjusted R-squared:  0.6048
## F-statistic: 158.8 on 8 and 817 DF,  p-value: < 2.2e-16
```

Tutte le variabili indipendenti risultano molto significative, tranne *Coarse.Aggregate* e *Fine.Aggregate*, che lo sono poco. L'R^2 e l'R^2_{adj} risultano intorno al valore 0.60, e reputiamo questo valore al limite della sufficienza. Ciò significa che il modello spiega il 60% della variabilità della variabile risposta.

Inoltre, il p-value del test $H0$: $i\ coefficienti\ sono\ tutti\ pari\ a\ zero$ vs. $H1$: $almeno\ un\ coefficiente$ è $diverso\ da\ zero$ è molto piccolo, stando a significare che almeno un coefficiente per i predittori considerati è significatamente diverso da zero.

Vediamo ore le diagnostiche grafiche dei residui del modello.



Il primo grafico, *Residuals vs Fitted*, rappresenta il diagramma a dispersione dei residui verso i valori predetti. Questi ultimi sembrano essere disposti in maniera casuale intorno alla retta x=0 e non sono quindi presenti pattern evidenti, anche se si nota un addensamento dei punti per i valori fittati minori di 40.

Il secondo grafico, il diagramma quantile-quantile dei residui standardizzati, ci suggerisce che è ragionevole pensare che la distribuzione degli errori del modello sia di tipo normale. Infatti, gli scostamenti dei valori dalla bisettrice del 1° e 3° quadrante non sembrano particolarmente importanti.

Il terzo grafico, *Scale Location*, individua una situazione di possibile eteroschedasticità: si nota infatti anche in questo grafico un addensamento dei punti per i valori più bassi dei Fitted Values.

Tramite il test di Breusch-Pagan si può verificare formalmente l'ipotesi di omoschedasticità dei residui del modello. Le ipotesi del test sono: $H0$: $la\ varianza\ degli\ errori$ è $costante$ vs $H1$: $la\ varianza\ degli\ errori\ non$ è $costante$.

```
##
##  studentized Breusch-Pagan test
##
## data:  mod1
## BP = 101.19, df = 8, p-value < 2.2e-16
```

Il basso p-value per questo test mostra che effettivamente non ci si trova in una situazione di omoschedasticità dei residui.

L'ultimo grafico, *Residuals vs Leverage*, mostra le Distanze di Cook; è possibile notare quali sono le osservazioni che hanno una particolare influenza sul modello di regressione. In questo caso, esse si tratta delle osservazioni 77, 611 e 225. Poiché i valori influenti pesano nella stima dei coefficienti del modello, decidiamo di eliminarli e di fittare nuovamente lo stesso modello per osservare se vi sono miglioramenti.

Procediamo dunque con la rimozione dei valori che hanno distanza di Cook maggiore della soglia consigliata in letteratura $4/n$ dove n è il numero di osservazioni.

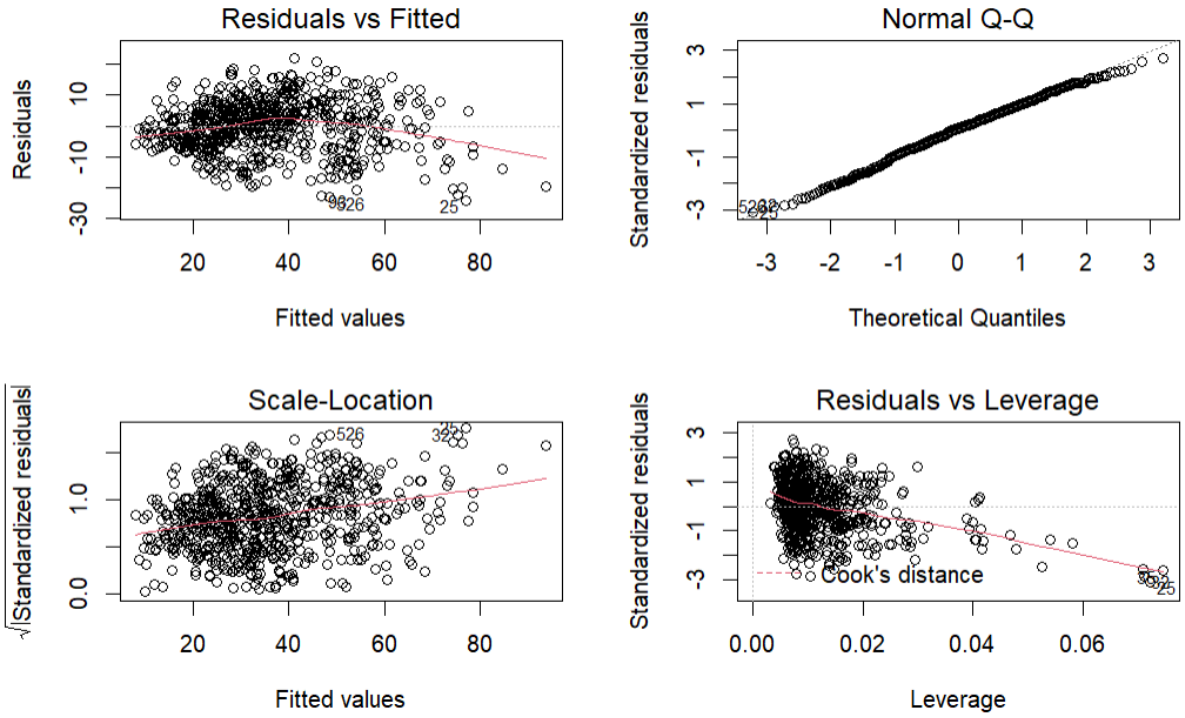
```
## [1] 758   9
```

La dimensione del training set è passata da 826 a 758, quindi sono state eliminate 68 osservazioni influenti.

Vediamo il nuovo fitting del modello completo.

```
##
## Call:
## lm(formula = Strength ~ ., data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.4349  -5.2882   0.7146   5.4844  22.0681
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -32.933737  25.484301  -1.292 0.196647
## Cement               0.130434   0.007927  16.454  < 2e-16 ***
## Blast.Furnace.Slag   0.116818   0.009665  12.087  < 2e-16 ***
## Fly.Ash              0.095483   0.011778   8.107 2.11e-15 ***
## Water               -0.179470   0.038958  -4.607 4.81e-06 ***
## Superplasticizer     0.310824   0.091893   3.382 0.000756 ***
## Coarse.Aggregate     0.021151   0.008927   2.369 0.018071 *
## Fine.Aggregate       0.028360   0.010270   2.761 0.005895 **
## Age                  0.171311   0.006647  25.771  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.207 on 749 degrees of freedom
## Multiple R-squared:  0.7565, Adjusted R-squared:  0.7539
## F-statistic: 290.9 on 8 and 749 DF,  p-value: < 2.2e-16
```

Notiamo che il valore di *Radj2* è notevolmente migliorato arrivando a 0.75. Inoltre,la significatività delle variabili Coarse.Aggregate e Fine.Aggregate è migliorata, mentre quella dell'intercetta è rimasta molto bassa. Osserviamo i plot di diagnostica per questo modello.



Anche in questo caso il *Normal Q-Q Plot* suggerisce una distribuzione gaussiana degli errori, mentre lo *Scale-Location* sembra essere migliorato poiché si nota meno l'addensamento vicibile nelle diagnostiche del modello precedente, mentre nel *Residual vs Fitted* si nota ancora lo stesso addensamento di punti per i valori fittati minori di 40 e nel *Residual vs Leverage* si notano ancora dei punti distanti dalla nuvola, tuttavia evidentemente si tratta di punti con Leverage elevato ma non superiore alla soglia.

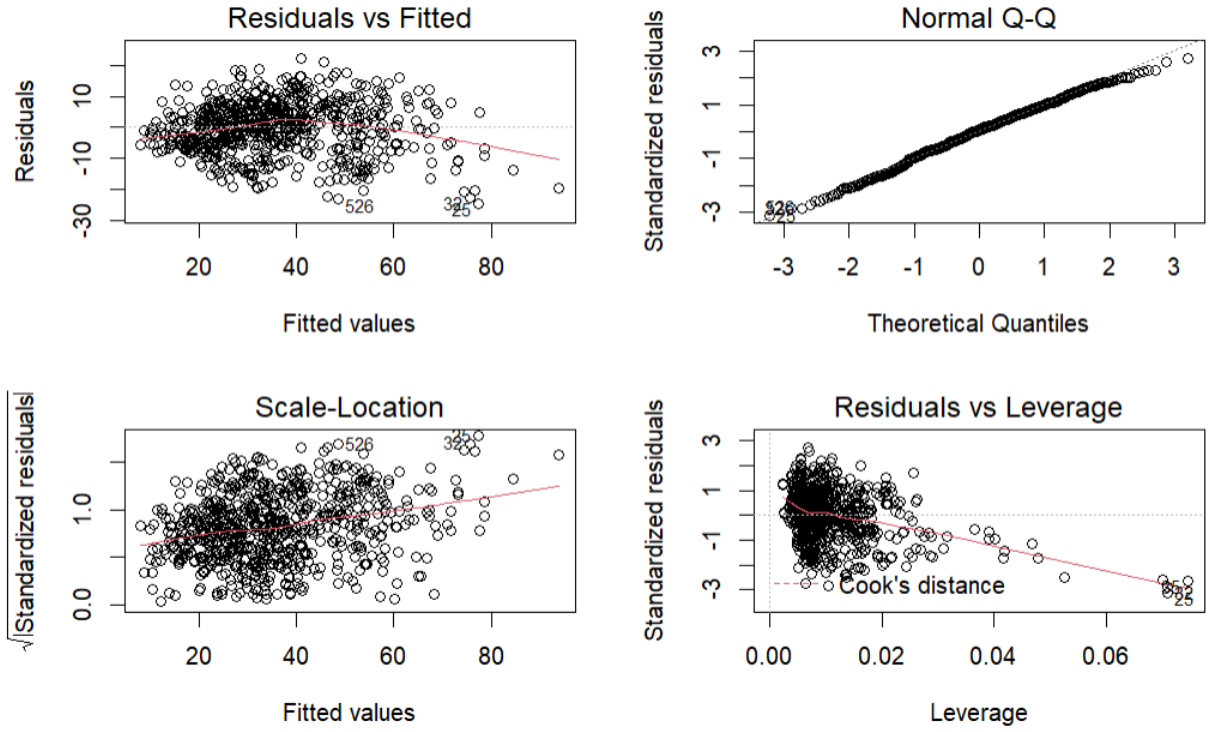
Fittiamo ora un modello che non comprende l'intercetta, poiché questa nell'ultimo modello risultava non significativa.

```
##
## Call:
## lm(formula = Strength ~ . - 1, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -24.8191  -5.3909   0.5849   5.5971  22.2623
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## Cement             0.121580   0.003989  30.481  < 2e-16 ***
## Blast.Furnace.Slag 0.106040   0.004886  21.705  < 2e-16 ***
## Fly.Ash            0.083700   0.007459  11.221  < 2e-16 ***
## Water             -0.225603   0.015608 -14.454  < 2e-16 ***
## Superplasticizer   0.269091   0.086070   3.126 0.001838 **
```

```
## Coarse.Aggregate    0.010204   0.002817   3.623 0.000311 ***
## Fine.Aggregate      0.015808   0.003337   4.738 2.59e-06 ***
## Age                 0.171166   0.006649  25.742  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.211 on 750 degrees of freedom
## Multiple R-squared:  0.9565, Adjusted R-squared:  0.956
## F-statistic:  2061 on 8 and 750 DF,  p-value: < 2.2e-16
```

Il modello ottenuto sembra essere molto soddisfacente: tutti i parametri sono molto significativi quello che lo è meno è *Superplasticizer* ad un livello di significatività dello 0.002), l'R^2_{adj} raggiunge un valore molto elevato, pari a 0.956.

Visualizziamo i plot di diagnostica dei residui:



La situazione dei residui non si è modificata rispetto al modello che comprendeva anche l'intercetta.

Scelta del modello migliore

Criterio di Akaike

Per i risultati ottenuti il modello migliore tra quelli fittati da noi sembra essere l'ultimo, il *mod3*, tuttavia decidiamo di utilizzare il metodo di Akaike per verificare formalmente ciò che abbiamo ipotizzato osservando le statistiche di adattamento, i t-test per i coefficienti e i grafici dei residui.

```
##
## Model selection based on AICc:
##
##       K    AICc Delta_AICc AICcWt Cum.Wt       LL
## mod3  9 5353.23       0.00   0.55   0.55 -2667.50
## mod2 10 5353.60       0.37   0.45   1.00 -2666.65
## mod1 10 6237.60     884.37   0.00   1.00 -3108.67
```

- K rappresenta il numero di parametri nel modello;
- AICc è l'information score del modello. La *c* minuscola indica che il valore è stato calcolato con un test AIC corretto per campioni di piccole dimensioni. Minore è questo valore, migliore è il modello;
- Delta_AICc è la differenza tra lo score di ciascun modello con quello del modello migliore;
- AICcWt: AICc weight, che è la proporzione relativa al modello preso in esame del potere predittivo totale fornito dal set completo di modelli
- Cum.Wt è la somma cumulata dei pesi AICc;
- LL è la Log-likelihood. Questo valore dà un'indicazione circa quanto il modello sia verosimile, dati i dati.

In effetti il terzo modello risulta, per i parametri considerati dal metodo di Akaike, solo leggermente migliore del secondo. Considerando anche il miglior indice di adattamento R^2_{adj} scegliamo di considerare il terzo modello come quello migliore.

Partendo dal modello individuato, utilizziamo la tecnica dello stepAIC per verificare se ci sono modelli che noi non abbiamo considerato migliori del nostro nella spiegazione dell'andamento di *Strength*. Applichiamo dunque il processo con l'opzione $direction = 'both'$, che lo rende una procedura a passi basata su aggiunte o eliminazioni di variabili seguendo il criterio di informazione di Akaike (AIC).

```
## Start:  AIC=3199.88
## Strength ~ (Cement + Blast.Furnace.Slag + Fly.Ash + Water + Superplasticizer +
##     Coarse.Aggregate + Fine.Aggregate + Age) - 1
##
##                      Df Sum of Sq    RSS    AIC
## <none>                             50567 3199.9
## - Superplasticizer    1       659  51226 3207.7
## - Coarse.Aggregate    1       885  51452 3211.0
## - Fine.Aggregate      1      1513  52080 3220.2
## - Fly.Ash             1      8489  59056 3315.5
## - Water               1     14087  64654 3384.2
## - Blast.Furnace.Slag  1     31762  82329 3567.3
## - Age                 1     44676  95243 3677.8
## - Cement              1     62641 113208 3808.8
```

In questo caso vediamo che il modello che noi abbiamo fittato, fit3, anche secondo il metodo di Akaike è il migliore, poiché questo non apporterebbe alcuna modifica per migliorarlo ulteriormente.

Confronto di modelli annidati tramite Anova

Un'altro metodo che si può utilizzare per il confronto di modelli è il test Anova. Possiamo utilizzarlo in questo caso perché uno dei due modelli presenta una selezione di regressori dell'altro, perché si tratta cioè di modelli annidati. Possiamo confrontare tra loro con questo metodo solo il *mod2* e il *mod3*, poiché il *mod1* non è stato fittato sugli stessi dati (il dataset conteneva ancora i valori anomali).

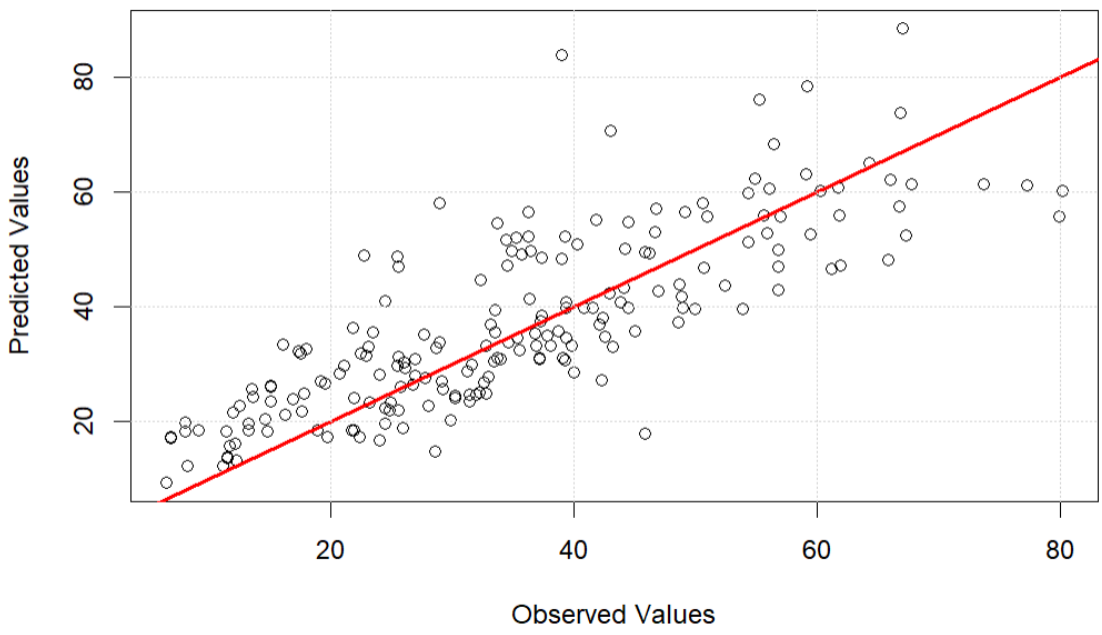
```
## Analysis of Variance Table
##
## Model 1: Strength ~ Cement + Blast.Furnace.Slag + Fly.Ash + Water +
Superplasticizer +
##     Coarse.Aggregate + Fine.Aggregate + Age
## Model 2: Strength ~ (Cement + Blast.Furnace.Slag + Fly.Ash + Water +
Superplasticizer +
##     Coarse.Aggregate + Fine.Aggregate + Age) - 1
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    749 50454
## 2    750 50567 -1    -112.5 1.6701 0.1966
```

Il test Anova mostra che non ci sono evidenze empiriche per rifiutare con un'opportuna significatività l'ipotesi che il *mod2* sia migliore del *mod3*. Tuttavia il modello che presenta il miglior adattamento valutato tramite R^2_{adj} è il terzo, che per questo potremmo considerare leggermente migliore.

Prediction sul test set

Con la funzione *predict* utilizziamo il terzo modello per stimare i valori del target del test set e uniamo la colonna dei valori previsti al dataframe di test.

Plottiamo infine i valori osservati e i valori predetti dal modello da noi stimato:



Questo plot mostra che il modello sembra predire abbastanza bene i valori della variabile target *Strength*.

Un campione di osservazioni del test set è riportato di seguito:

```
##     Cement Blast.Furnace.Slag Fly.Ash Water Superplasticizer Coarse.Aggregate
## 45   427.5               47.5     0.0 228.0              0.0              932
## 52   190.0              190.0     0.0 228.0              0.0              932
## 53   237.5              237.5     0.0 228.0              0.0              932
## 236  213.8               98.1    24.5 181.7              6.7             1066
## 318  251.8                0.0    99.9 146.1             12.4             1006
##     Fine.Aggregate Age Strength predicted
## 45           594.0  90    41.54  39.87893
## 52           670.0 180    46.93  42.72078
## 53           594.0  90    33.12  36.92635
## 236          785.5  14    17.84  24.94779
## 318          899.8  56    44.14  43.42519
```

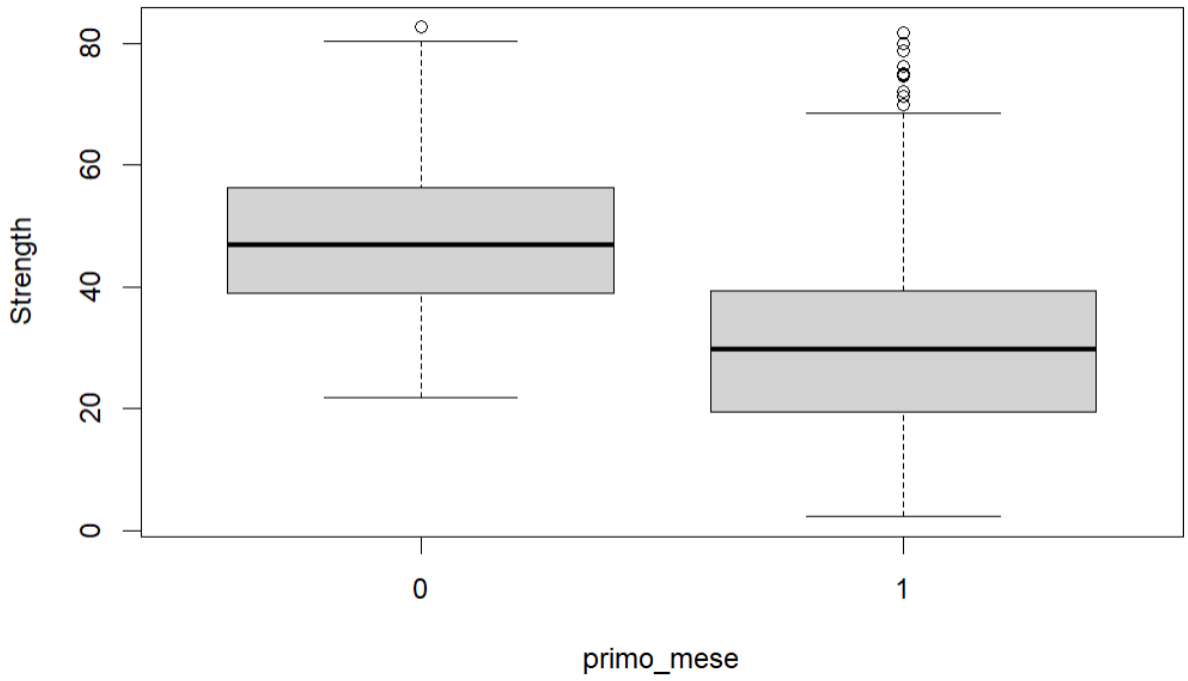
Curiosità: la media di *Strength* nei primi 30 giorni dalla posa del calcestruzzo è diversa dalla media di *Strength* nel periodo successivo?

Per concludere, applichiamo un test d'ipotesi per rispondere alla domanda che ci siamo poste. Considerando il dataset completo *Concrete*, dividiamo la variabile *Age* in due gruppi che comprendono le osservazioni dei primi 30 giorni e quelle dei giorni successivi. Il sistema di ipotesi è:

$H0$: *la media di Strength nei due gruppi* è *uguale* vs
$H1$: *la media di Strength nei due gruppi* è *diversa*.

Definiamo una nuova variabile binaria *primo_mese* che presenta valore 1 se Age è minore o uguale a 30 e 0 altrimenti.

Plottiamo la variabile *Strength* stratificata per *primo_mese*:



Applichiamo il test t per due campioni, ipotizzando che la varianza tra i due campioni sia la stessa visto che le osservazioni sono state estratte dalla medesima variabile casuale:

```
t.test(concrete$Strength ~ concrete$primo_mese, var.equal=TRUE, conf.level=0.95)

##
##  Two Sample t-test
##
## data:  concrete$Strength by concrete$primo_mese
## t = 16.962, df = 1028, p-value < 2.2e-16
## alternative hypothesis: true difference in means between group 0 and group 1 is
not equal to 0
## 95 percent confidence interval:
##  15.50231 19.55841
## sample estimates:
## mean in group 0 mean in group 1
##        48.56577        31.03541
```

Il valore molto piccolo di p-value suggerisce che la differenza delle medie dei due gruppi sia significativa.