Bike sharing a Londra

Cumitini Aldo
Ferioli Stefano
Boccia Flavio

Presentazione del dataset

date	season	temp	feeltemp	humidity	wind_speed	weather	is_holiday	is_weekend	is_restday	total
04/01/2015	winter	2.5	0.6	94.3	7.5	cloudy	0	1	1	9234
05/01/2015	winter	8	6.7	80.3	8.9	cloudy	0	0	0	20372
06/01/2015	winter	7.9	5.3	78.9	16	clear	0	0	0	20613
07/01/2015	winter	7.5	4.5	78.1	19.8	clear	0	0	0	21064
08/01/2015	winter	9.8	7.8	79.3	20.5	rain	0	0	0	15601
09/01/2015	winter	12.7	12.3	74.9	32.9	cloudy	0	0	0	22104
10/01/2015	winter	10.5	8.7	66.1	34.3	cloudy	0	1	1	14709
11/01/2015	winter	6.6	2.5	67.6	26.6	clear	0	1	1	14575
12/01/2015	winter	11.1	9.8	76.6	28.2	rain	0	0	0	17199
13/01/2015	winter	8.6	6.1	75.8	21.2	rain	0	0	0	24697
14/01/2015	winter	6.5	2.5	67.1	25.8	clear	0	0	0	23565

728 misurazioni

```
    date              season         temp            feeltemp          humidity        wind_speed       weather      is_holiday         is_weekend        is_restday         total
Length:728         winter:177   Min.   : 1.30   Min.   :-2.400   Min.   :46.90   Min.   : 2.80   clear :438   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000   Min.   : 4869
Class :character   spring:186   1st Qu.: 8.60   1st Qu.: 6.475   1st Qu.:65.50   1st Qu.:11.20   cloudy:203   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:21922
Mode  :character   summer:185   Median :12.50   Median :12.400   Median :72.55   Median :15.20   rain  : 87   Median :0.00000   Median :0.0000   Median :0.0000   Median :26968
                   autumn:180   Mean   :12.47   Mean   :11.516   Mean   :72.35   Mean   :15.96                Mean   :0.02198   Mean   :0.2871   Mean   :0.3091   Mean   :27157
                                3rd Qu.:16.20   3rd Qu.:16.200   3rd Qu.:79.22   3rd Qu.:19.90                3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:33362
                                Max.   :27.40   Max.   :27.400   Max.   :98.70   Max.   :41.90                Max.   :1.00000   Max.   :1.0000   Max.   :1.0000   Max.   :46021
```

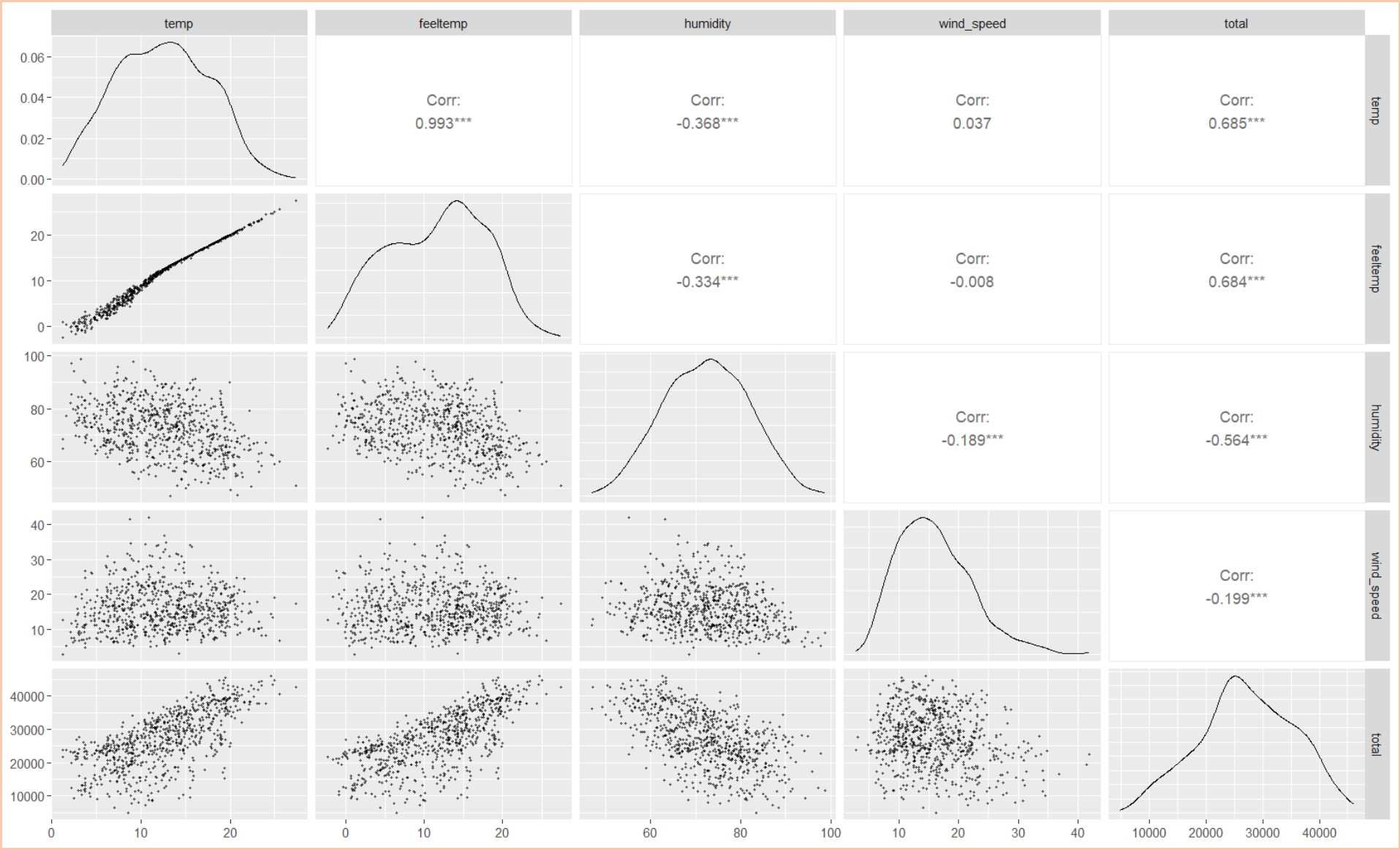
https://www.kaggle.com/datasets/hmavrodiev/london-bike-sharing-dataset

https://cycling.data.tfl.gov.uk/

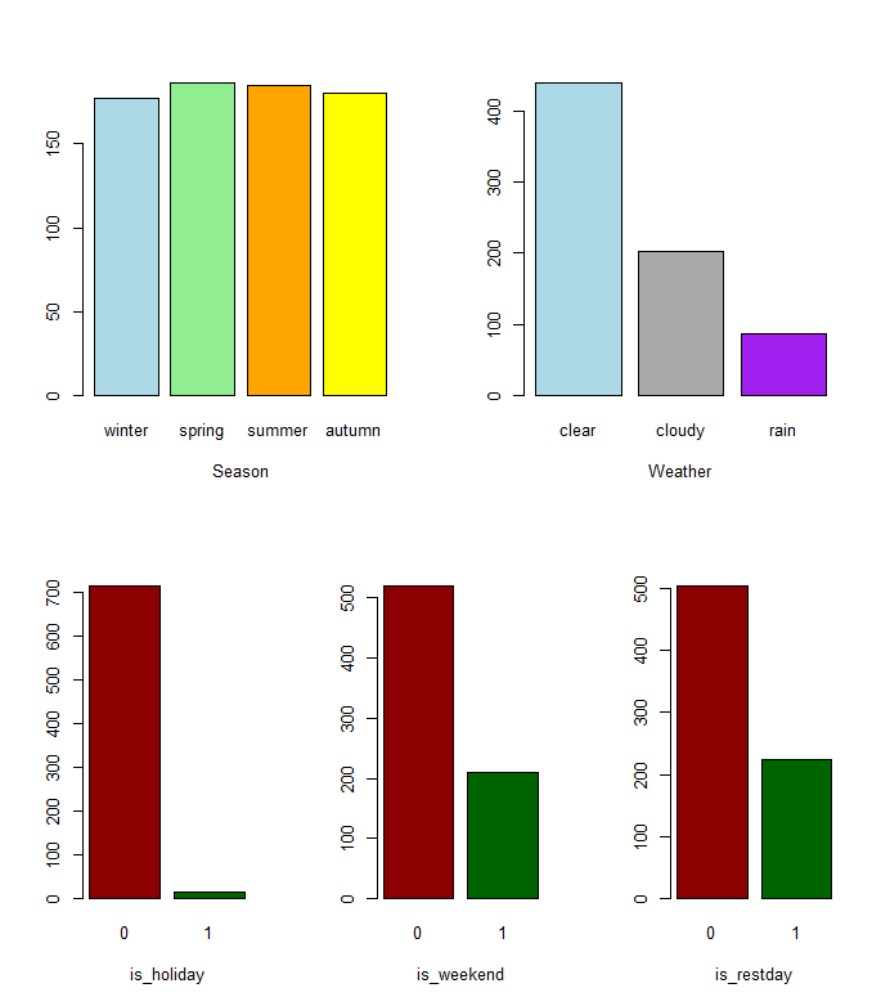
Struttura della presentazione

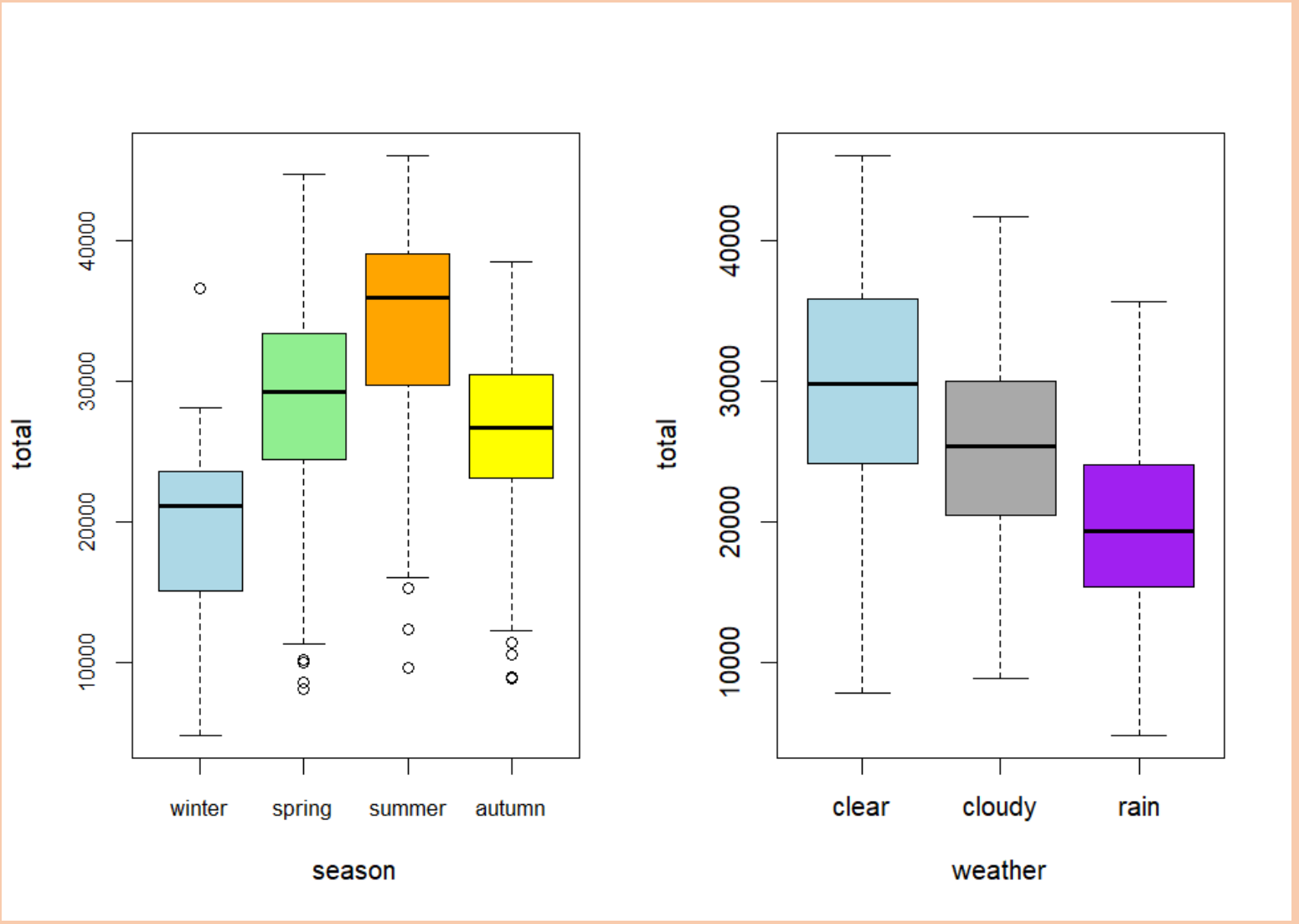
- Analisi qualitativa del dataset
- Costruzione di un modello lineare predittivo
- Verifica delle ipotesi di normalità ed eliminazione dei punti critici
- Verifica della bontà del modello
- Test ANOVA sul numero di noleggi tra le diverse stagioni

Analisi qualitativa del dataset



Analisi qualitativa del dataset





Modello lineare

Dividiamo il dataset in due parti in modo random: train-set con 600 osservazioni e test-set con 128 osservazioni.

Generiamo il primo modello lineare sul train-set

```
> g = lm(total ~ temp + humidity + wind_speed + is_restday + weather, data = train)
> summary(g)

Call:
lm(formula = total ~ temp + humidity + wind_speed + is_restday +
    weather, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-19247.8  -2317.3    127.3   2411.7  19202.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   45421.13    1897.55  23.937  < 2e-16 ***
temp            893.23      33.31  26.813  < 2e-16 ***
humidity       -296.08      22.25 -13.309  < 2e-16 ***
wind_speed     -352.69      26.53 -13.294  < 2e-16 ***
is_restday    -5398.56     346.49 -15.581  < 2e-16 ***
weathercloudy  -459.66     417.34  -1.101    0.271
weatherrain   -4699.37     566.37  -8.297 7.19e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3841 on 593 degrees of freedom
Multiple R-squared:  0.7817,           Adjusted R-squared:  0.7795
F-statistic: 353.9 on 6 and 593 DF,  p-value: < 2.2e-16
```

Notiamo che

- Il P-value dell’F-test è 2.2e-16, quindi c’è almeno una covariata significatica
- $R^2_{adj} = 0.7795$
- La covariata dummy weathercloudy è l’unica variabile non significativa
- Tutte le altre covariate sembrerebbero essere significative

Modello lineare

Generiamo quindi un altro modello lineare escludendo la covariata weathercloudy

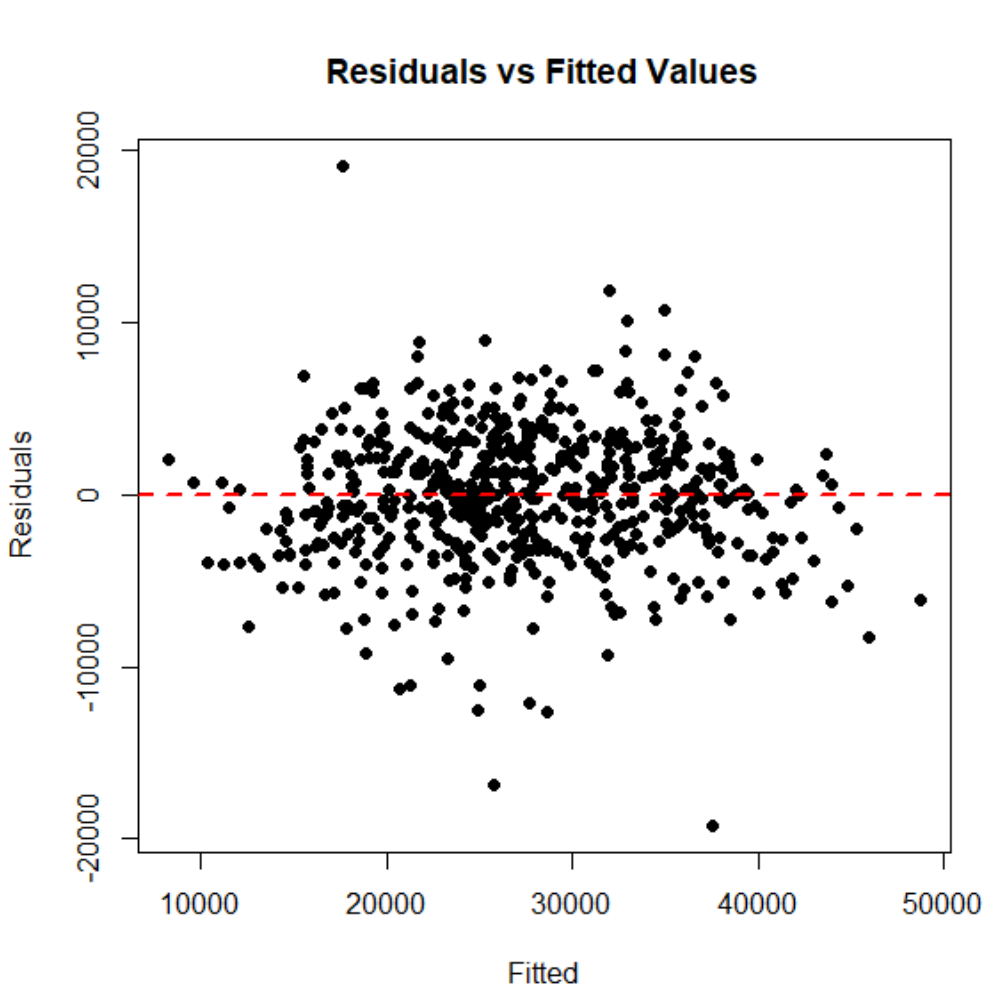
```
> train$weatherrain = ifelse(train$weather == "rain",1,0)
> g = lm(total ~ temp + humidity + wind_speed + is_restday + weatherrain, data = train)
> summary(g)

Call:
lm(formula = total ~ temp + humidity + wind_speed + is_restday +
    weatherrain, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-19203.8  -2293.3     80.4   2411.8  19059.5

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 46279.05    1730.61  26.741   <2e-16 ***
temp          888.66      33.06  26.880   <2e-16 ***
humidity     -308.20      19.34 -15.939   <2e-16 ***
wind_speed   -357.21      26.22 -13.626   <2e-16 ***
is_restday  -5427.98     345.52 -15.710   <2e-16 ***
weatherrain -4442.19     516.09  -8.607   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3841 on 594 degrees of freedom
Multiple R-squared:  0.7812,           Adjusted R-squared:  0.7794
F-statistic: 424.3 on 5 and 594 DF,  p-value: < 2.2e-16
```

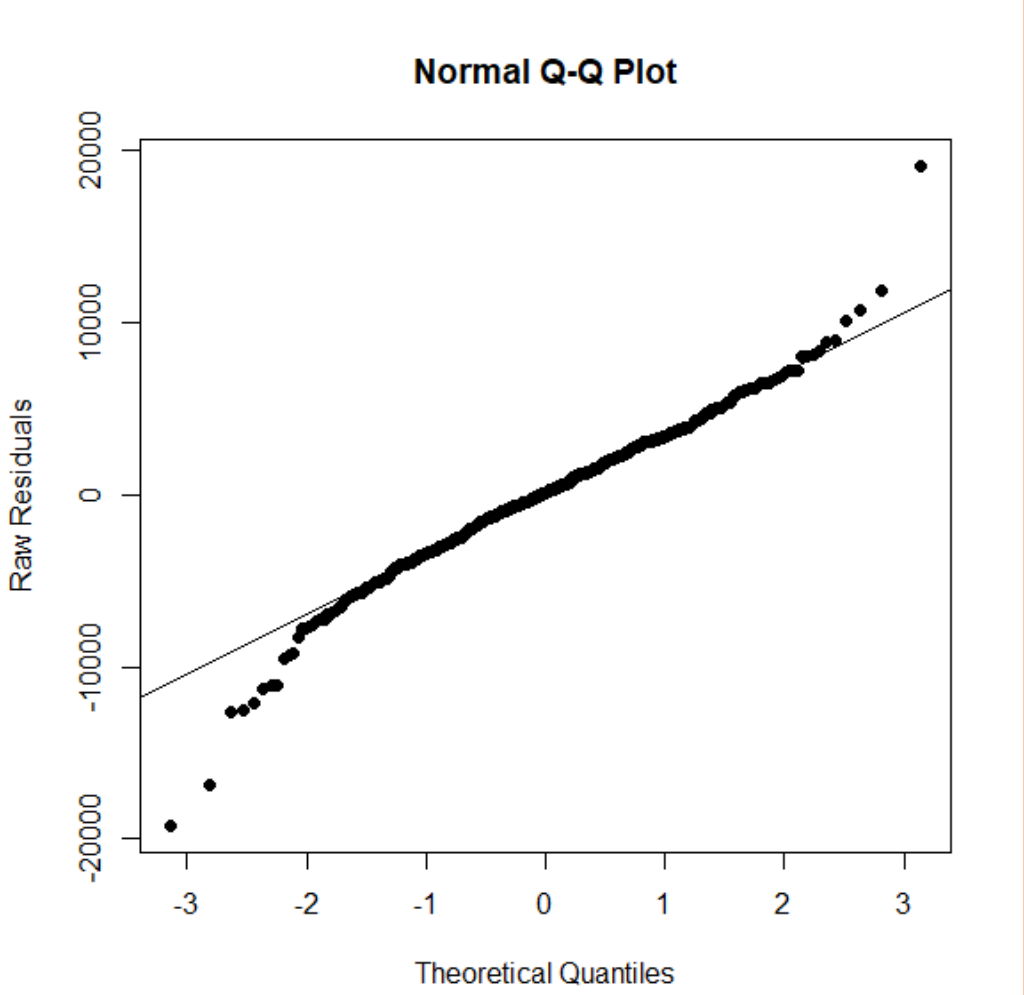
Notiamo che

- $R^2_{adj} = 0.7794$ (rimane praticamente uguale)
- Tutte le covariate rimanenti sono molto significative

Verifichiamo la normalità dei residui

Eseguendo lo Shapiro test sui residui del modello lineare si ottiene un p-value di 8.899e-09





Eliminiamo i punti critici

I residui standardizzati e i residui studentizzati coincidono

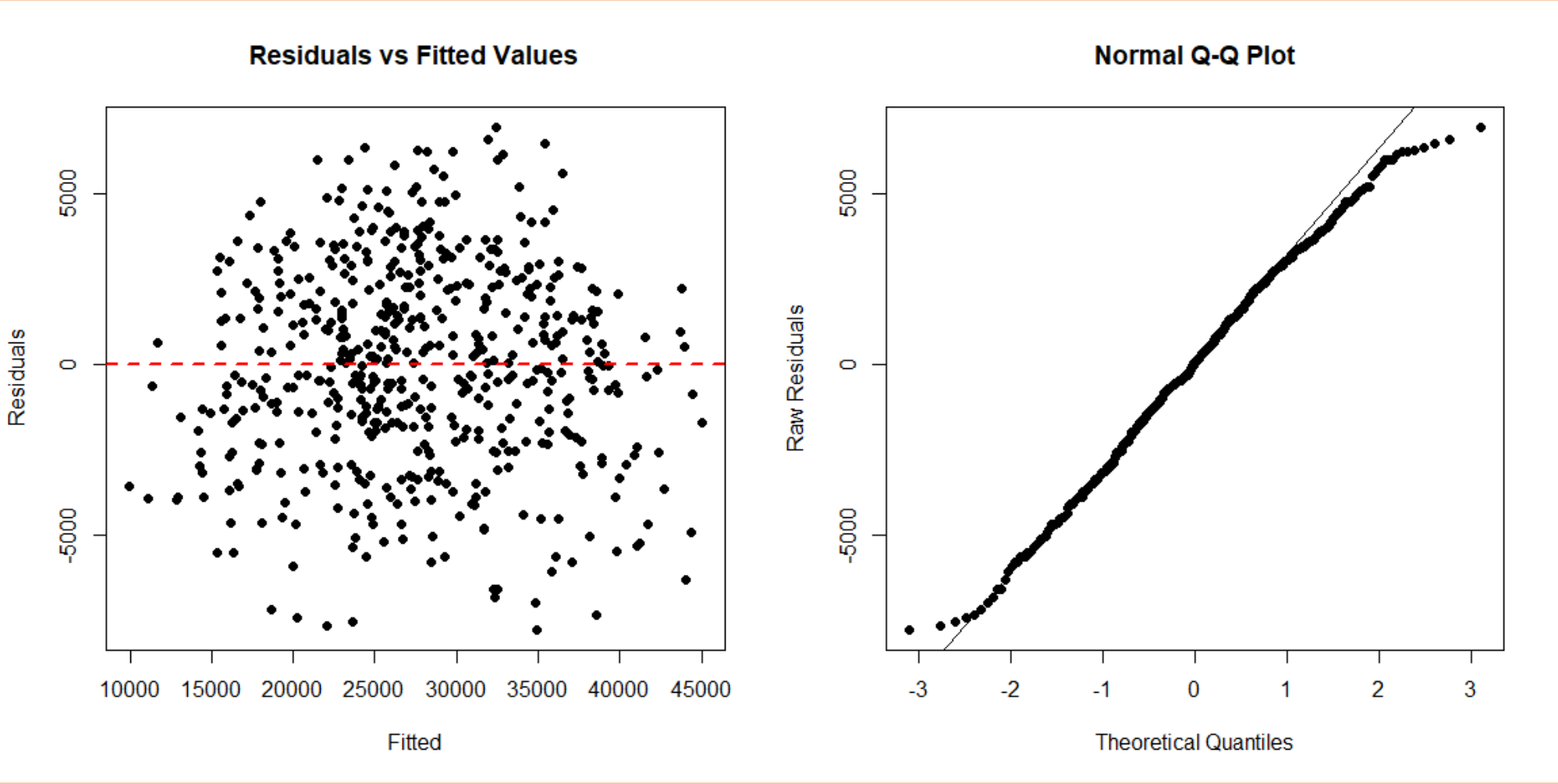
Eliminiamo i punti critici

Criteri utilizzati	AIC	R^2_{adj}
Dataset completo	11614.97	0.779
Punti leva	11333.14	0.772
Residui standardizzati	10923.21	0.840
Distanza di Cook	10611.24	0.842
Punti leva + residui standardizzati	10643.85	0.835
Residui standardizzati + distanza di Cook	10472.59	0.852
Punti leva + distanza di Cook	10193.02	0.836
Punti leva + residui standardizzati + distanza di Cook	10213.75	0.847

Il modello migliore risulta essere quello in cui sono stati eliminati i punti leva e i punti critici secondo la distanza di Cook

Verifichiamo di nuovo la normalità dei residui

Eliminando i punti leva e i punti critici secondo la distanza di Cook, rigeneriamo il modello e otteniamo un p-value per lo Shapiro test di 0.1387 quindi non rifiutiamo l'ipotesi di normalità.



Interpretazione del modello

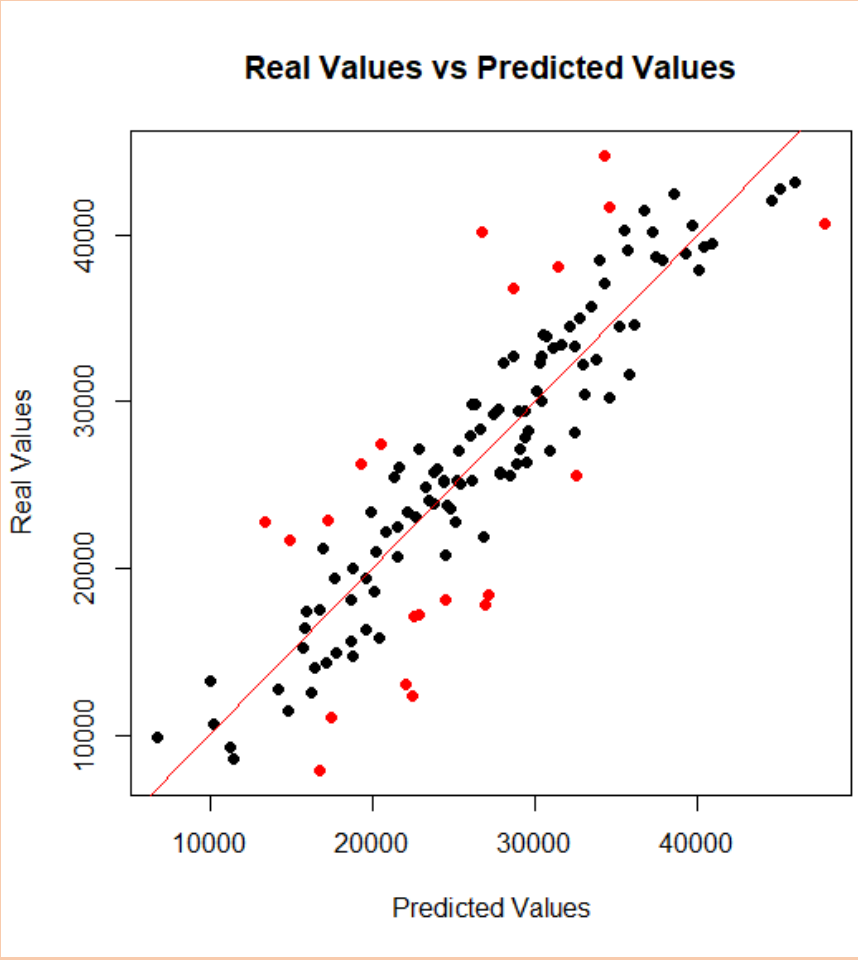
	Intercetta	temp	humidity	wind_speed	is_restday	weatherrain
Coefficienti β	43241	916.2	-275.5	-325.7	-5801.6	-4995.1
Coefficienti β trasformati	30503	21713	-14269	-10160.4	-5801.6	-4995.1

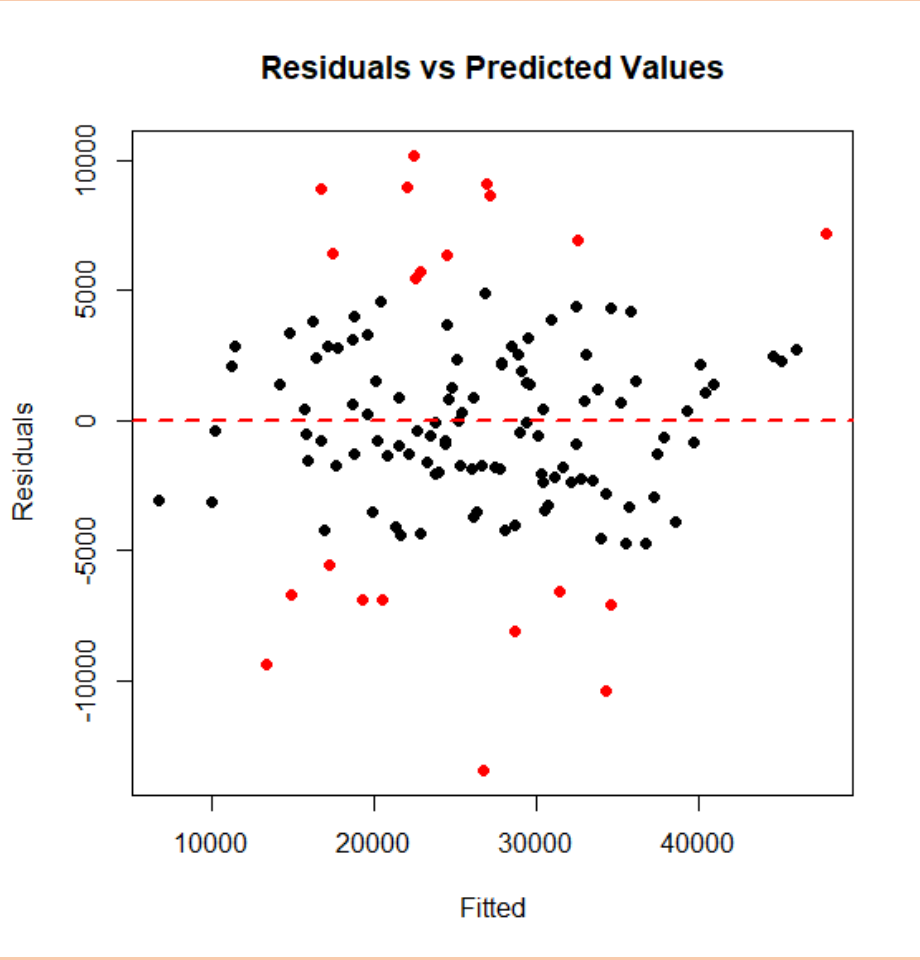
Trasformazione applicata alle covariate: $\frac{x - \min x}{\max x - \min x}$

Verifica della bontà del modello

Per verificare la bontà del nostro modello predittivo, lo abbiamo validato sul test-set che avevamo creato all'inizio.

I punti rossi rappresentano le misurazioni che cadono al di fuori dell'intervallo di confidenza al 90%.



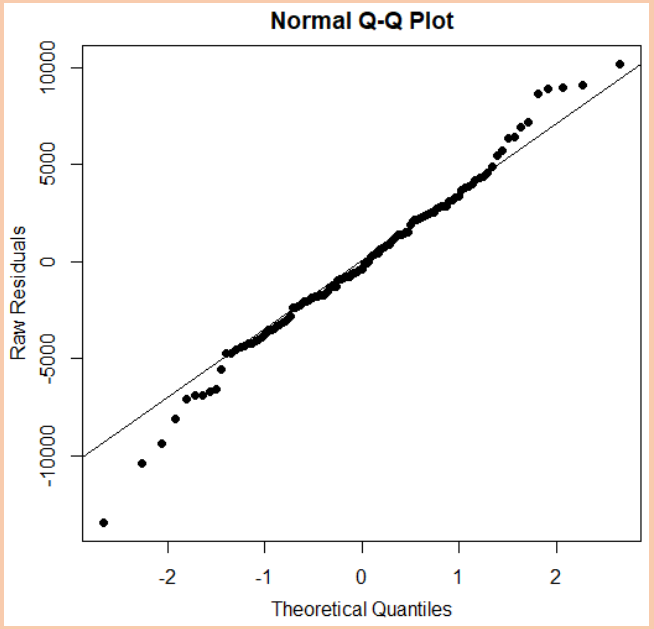


$R^2_{testset} = 0.8027$

$\rho_{real,pred} = 0.896$

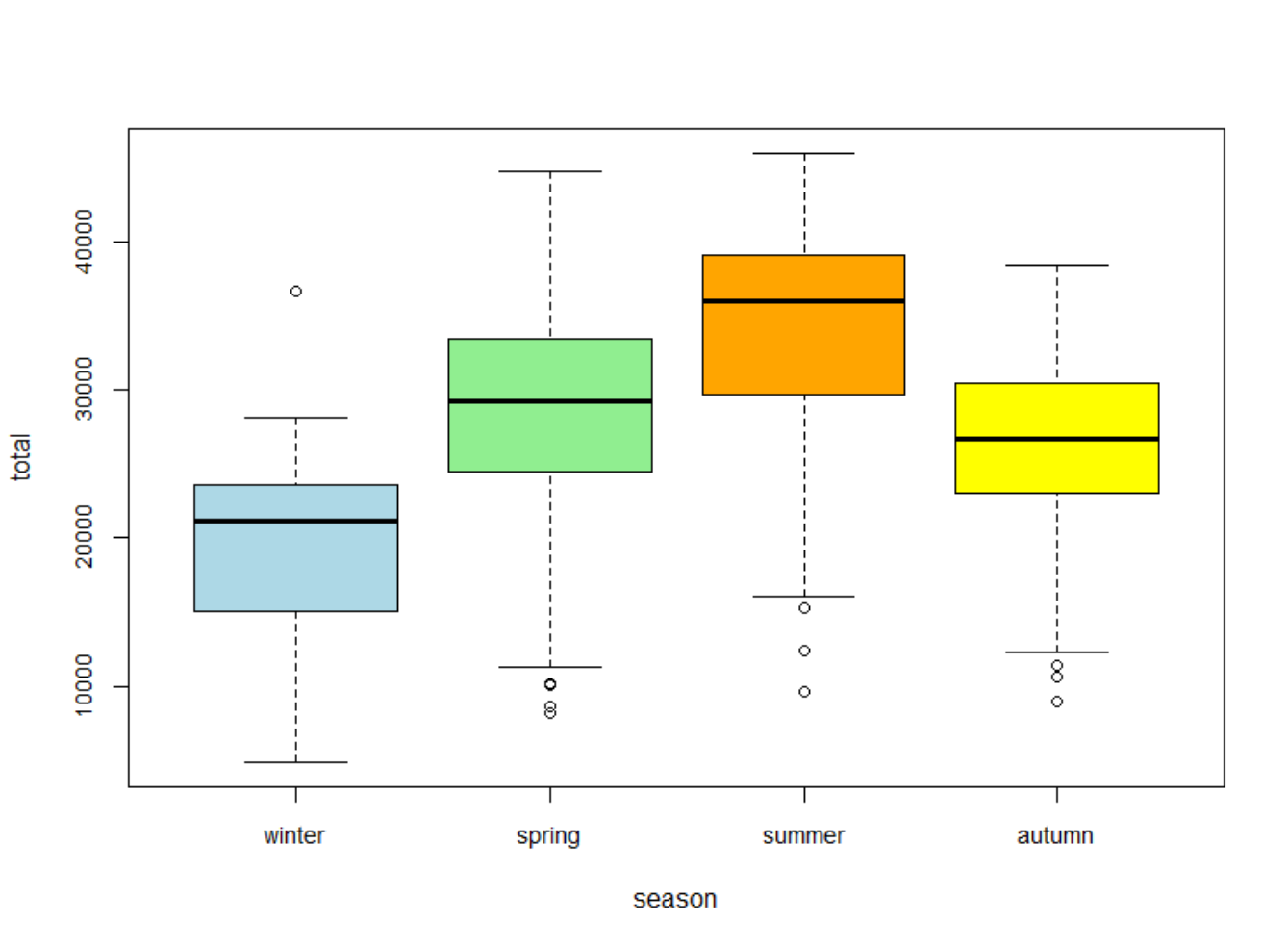
Percentuale dati nell'IC: 84%

P-value per lo Shapiro test: 0.3371

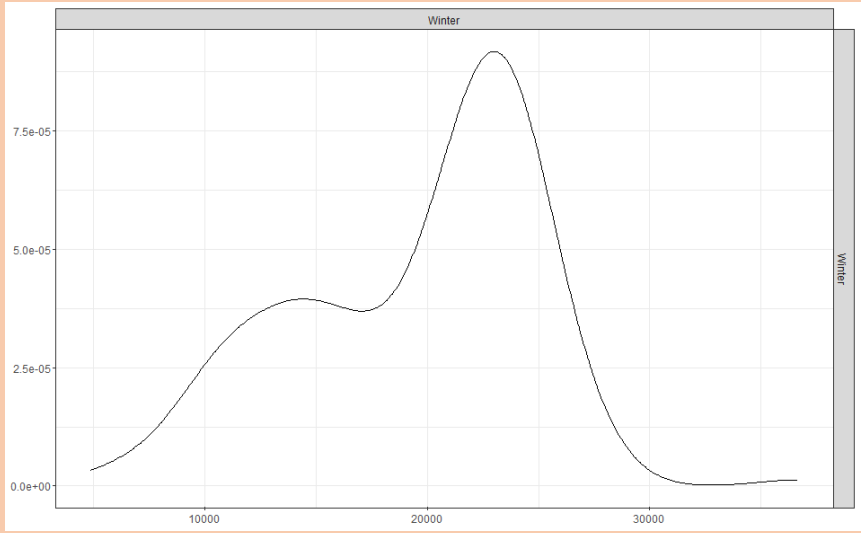


ANOVA

Osservando i boxplot divisi per stagione ci chiediamo se vi sia evidenza statistica per affermare che vi è differenza nella media delle bici noleggiate in ogni stagione. Lavoriamo sull'intero dataset.



Limitiamo la nostra analisi alle stagioni Spring, Summer e Autumn. La media per Winter è visibilmente inferiore rispetto alle altre. Inoltre, la distribuzione di total per winter non si presta a tale analisi.

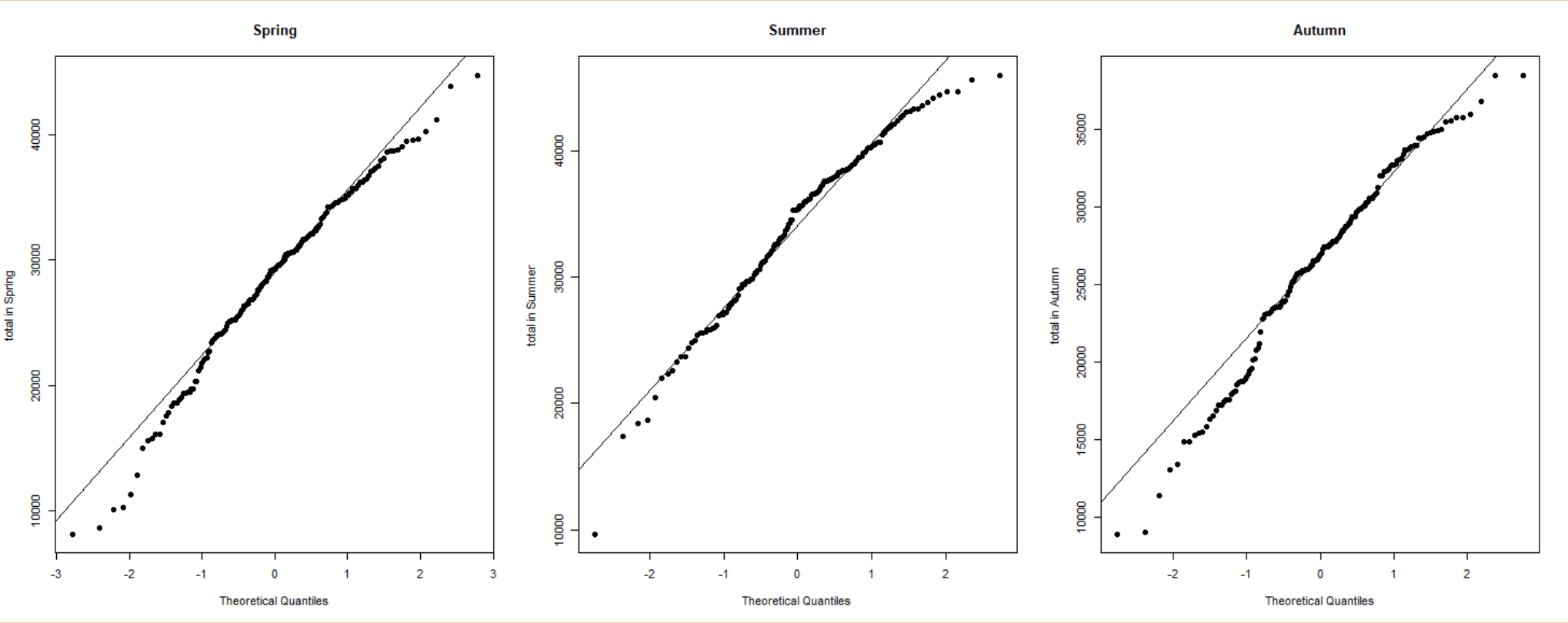


Controlliamo le ipotesi

Eseguendo lo Shapiro test sulla variabile total divisa per stagioni otteniamo

```
> tapply( dataset$total, dataset$season, function( x ) ( shapiro.test( x )$p ) )
      spring       summer       autumn
1.167488e-02 1.789999e-06 6.686083e-04
```

e quindi rifiutiamo l'ipotesi di normalità.



Trasformazione Box-Cox

Proviamo ad applicare una trasformazione Box-Cox per ottenere la normalità di total nelle diverse stagioni

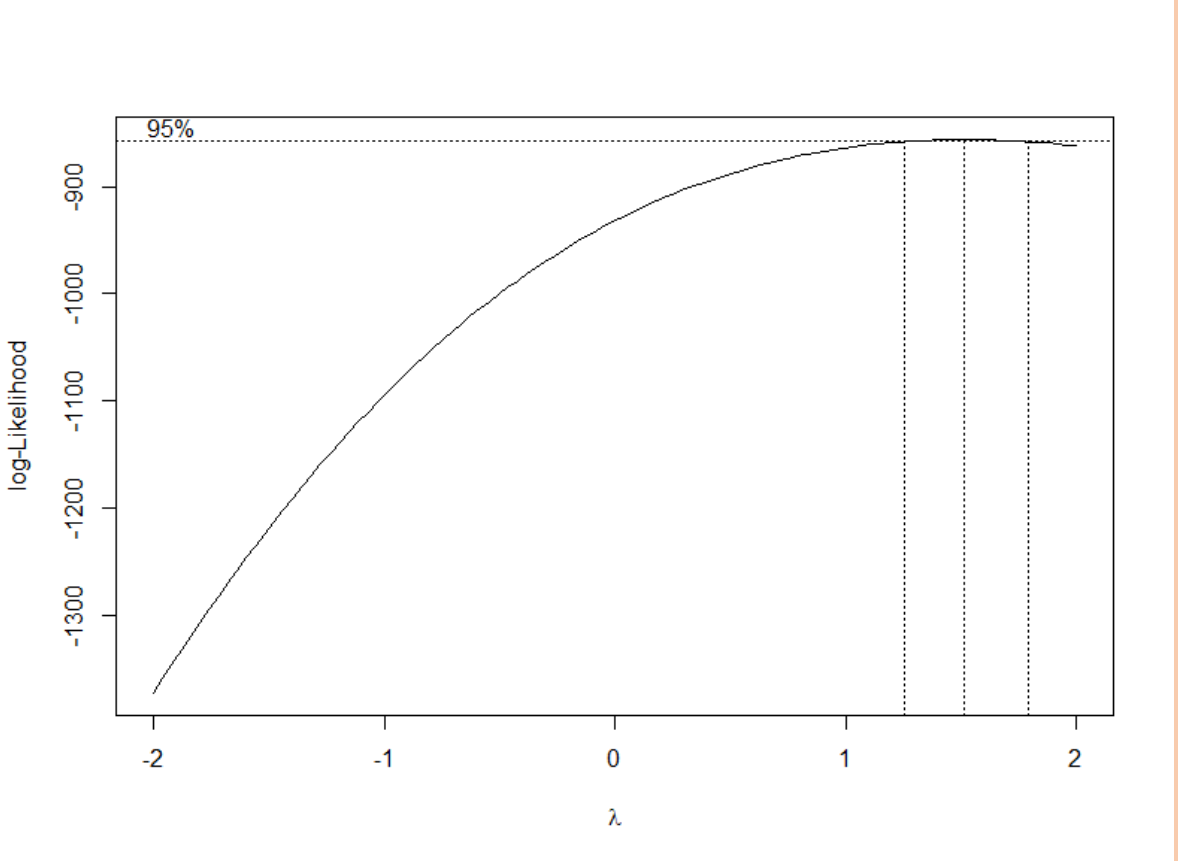
Otteniamo $\lambda = 1.52$

Eseguendo lo Shapiro test dopo la trasformazione otteniamo

```
> tapply( (dataset$total^best_lambda-1)/best_lambda,
dataset$season, function( x ) ( shapiro.test( x )$p ) )
spring     summer     autumn
0.81151870 0.07278747 0.10100644
```

Inoltre, eseguendo il Bartlett test e il Levene test otteniamo

```
bartlett.test( (dataset$total^best_lambda-1)/best_lambda,
dataset$season )

        Bartlett test of homogeneity of variances

data:  dataset$total and dataset$season
Bartlett's K-squared = 3.0646, df = 2, p-value = 0.216

leveneTest( (dataset$total^best_lambda-1)/best_lambda,
dataset$season )
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   2  0.9937 0.3709
      548
```



ANOVA

Dato che sono rispettate le ipotesi di normalità e omoschedasticità possiamo eseguire il test ANOVA sui gruppi

```
> summary(aov((total^best_lambda-1)/best_lambda ~ season, dataset))
             Df    Sum Sq   Mean Sq F value Pr(>F)
season        2 5.337e+14 2.668e+14   73.55 <2e-16 ***
Residuals   548 1.894e+15 3.628e+12
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Rifiutiamo l’ipotesi che le medie dei 3 gruppi siano uguali poiché il p-value è molto basso.

Infine, eseguiamo dei t-test per verificare che vi è differenza nella media delle bici noleggiate in ogni stagione

```
> t.test((datasetSp$total^best_lambda-1)/best_lambda, (datasetSu$total^best_lambda-1)/best_lambda, mu=0, alternative = "two.sided", paired = FALSE, var.equal = TRUE)

        Two Sample t-test

data:  (datasetSp$total^best_lambda - 1)/best_lambda and (datasetSu$total^best_lambda - 1)/best_lambda
t = -8.2155, df = 348, p-value = 4.206e-15
```

```
> t.test((datasetSu$total^best_lambda-1)/best_lambda, (datasetAu$total^best_lambda-1)/best_lambda, mu=0, alternative = "two.sided", paired = FALSE, var.equal = TRUE)

        Two Sample t-test

data:  (datasetSu$total^best_lambda - 1)/best_lambda and (datasetAu$total^best_lambda - 1)/best_lambda
t = 12.09, df = 337, p-value < 2.2e-16
```

```
> t.test((datasetAu$total^best_lambda-1)/best_lambda, (datasetSp$total^best_lambda-1)/best_lambda, mu=0, alternative = "two.sided", paired = FALSE, var.equal = TRUE)

        Two Sample t-test

data:  (datasetAu$total^best_lambda - 1)/best_lambda and (datasetSp$total^best_lambda - 1)/best_lambda
t = -3.4691, df = 359, p-value = 0.0005858
```