Bike sharing in London

Cumitini Aldo
Ferioli Stefano
Boccia Flavio

Overview of the 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 lines

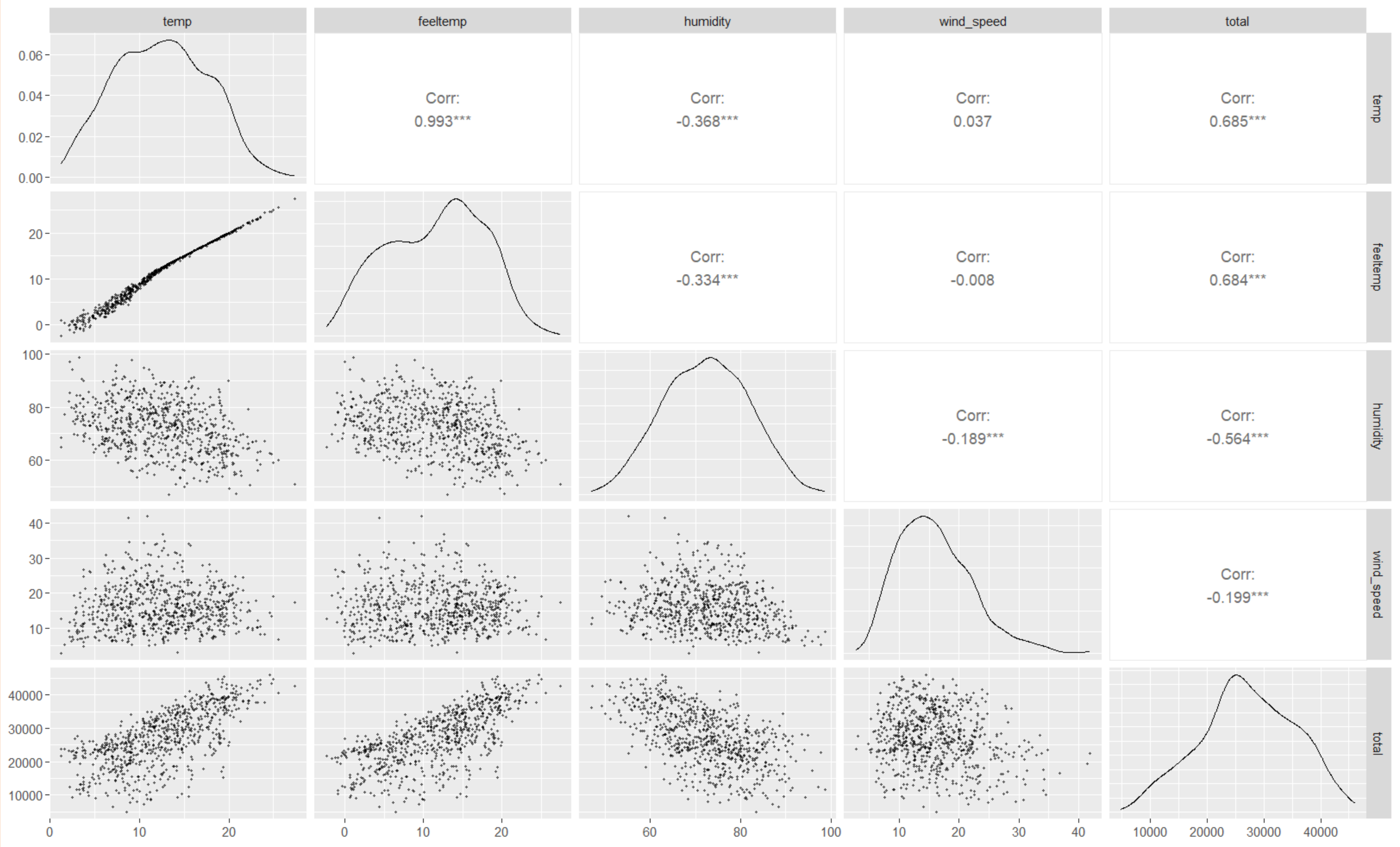
```
     date              season         temp          feeltemp          humidity        wind_speed       weather       is_holiday          is_weekend         is_restday          total
 Length:728         winter:177   Min.   : 1.30   Min.   :-2.400   Min.   :46.90   Min.   : 2.80   clear :438   Min.   :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/

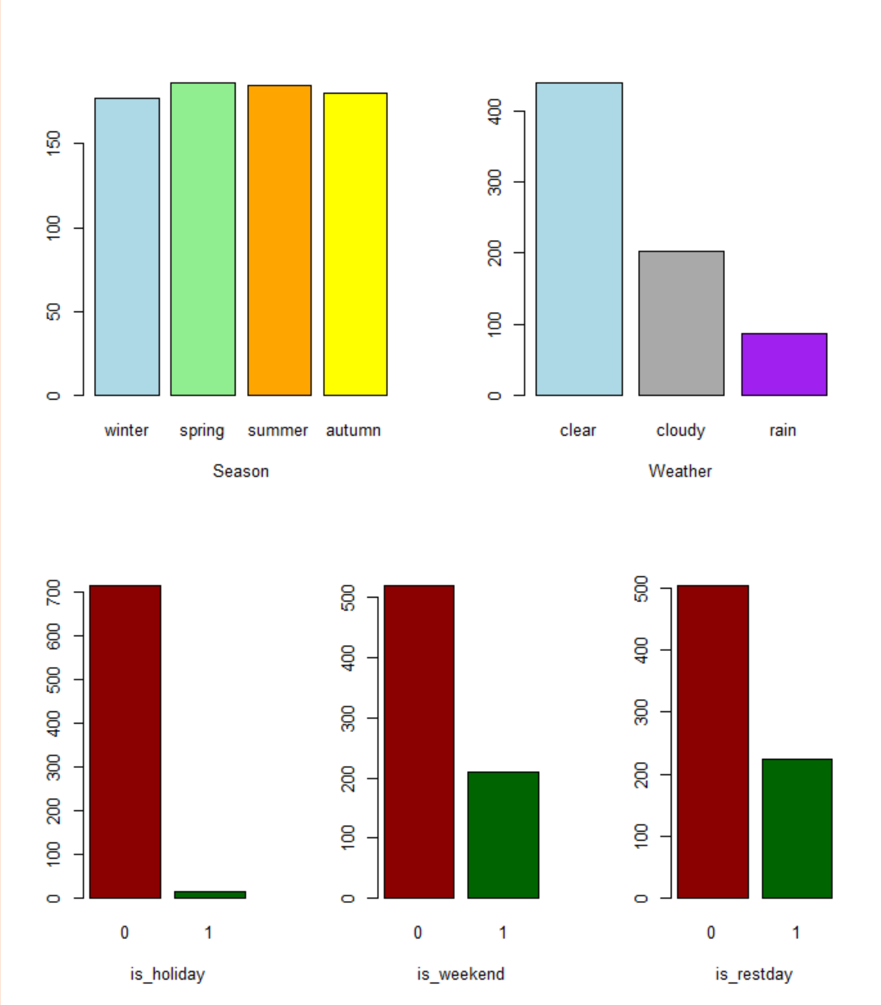
Presentation structure

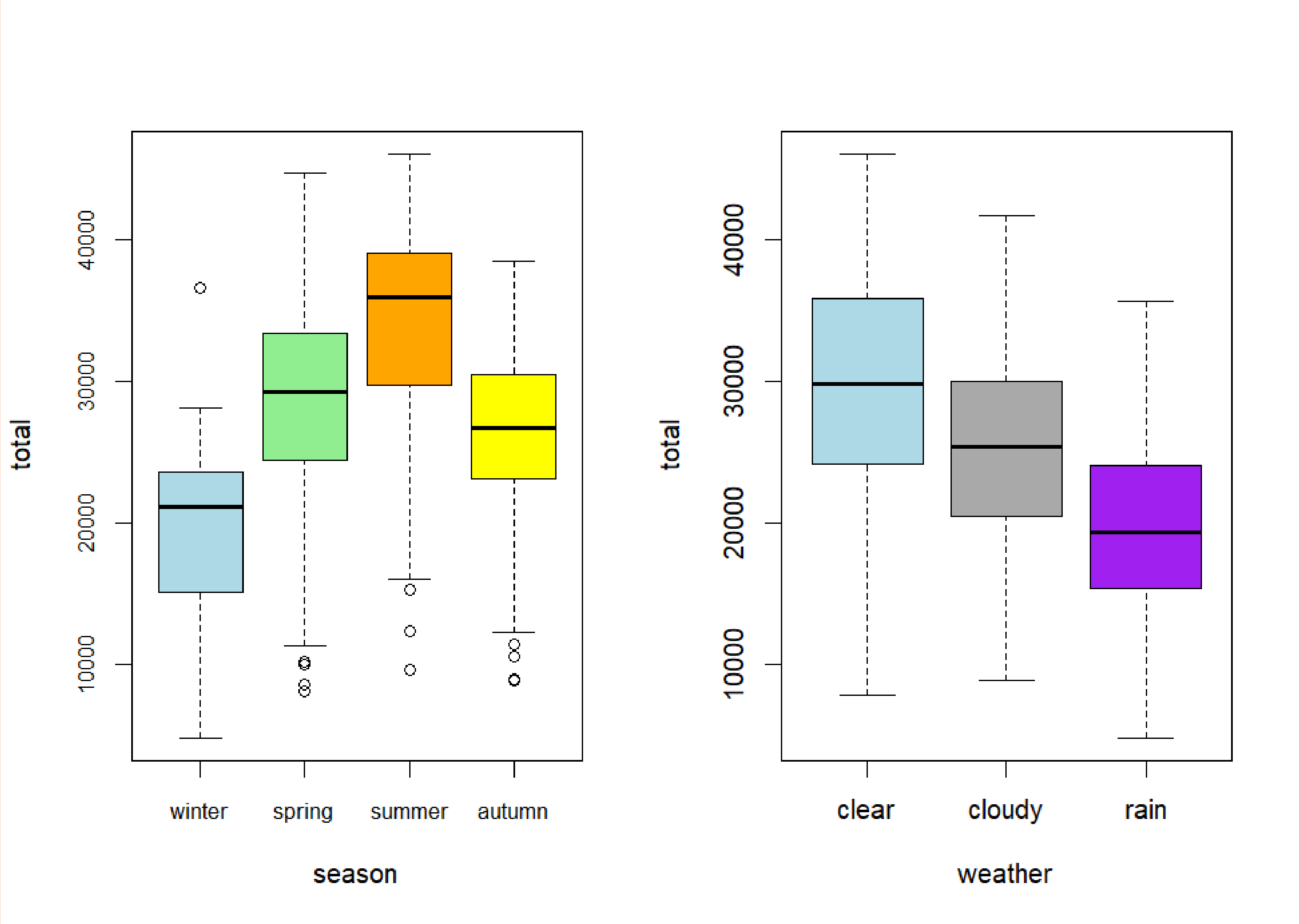
- Qualitative analysis of the dataset
- Construction of a predictive linear model
- Testing of normality assumptions and outliers' removal
- Assessment of the model performance
- ANOVA test on the number of rentals across different seasons

Qualitative analysis of the dataset



Qualitative analysis of the dataset





Linear model

We divide the dataset randomly in two parts: train-set with 600 lines and test-set with 128 lines.

We train the first linear model on the 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
```

We notice that

- The P-value of the F-test is 2.2e-16, indicating that there is at least one significant covariate.
- $R^2_{adj} = 0.7795$
- The dummy covariate weathercloudy is the only non-significant variable
- All other covariates appear to be significant

Linear model

We train another linear model removing the covariate weathercloudy

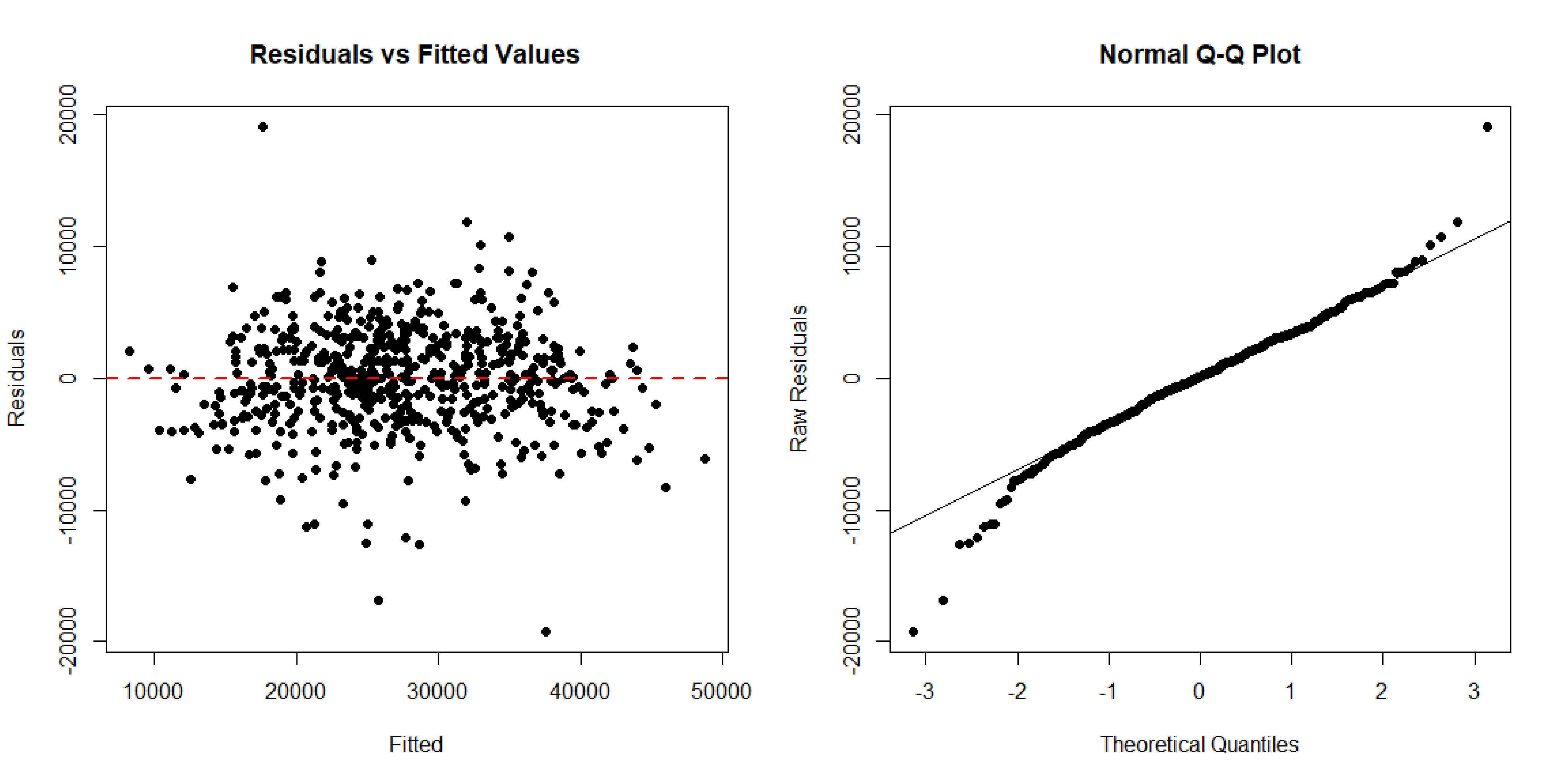
```
> train$weatherrain = ifelse(train$weather == "rain",1,0)
> g = lm(total ~ temp + humidity + wind_speed + is_restday + weatherrain, data = train)
> summary(g)

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
```

We notice that

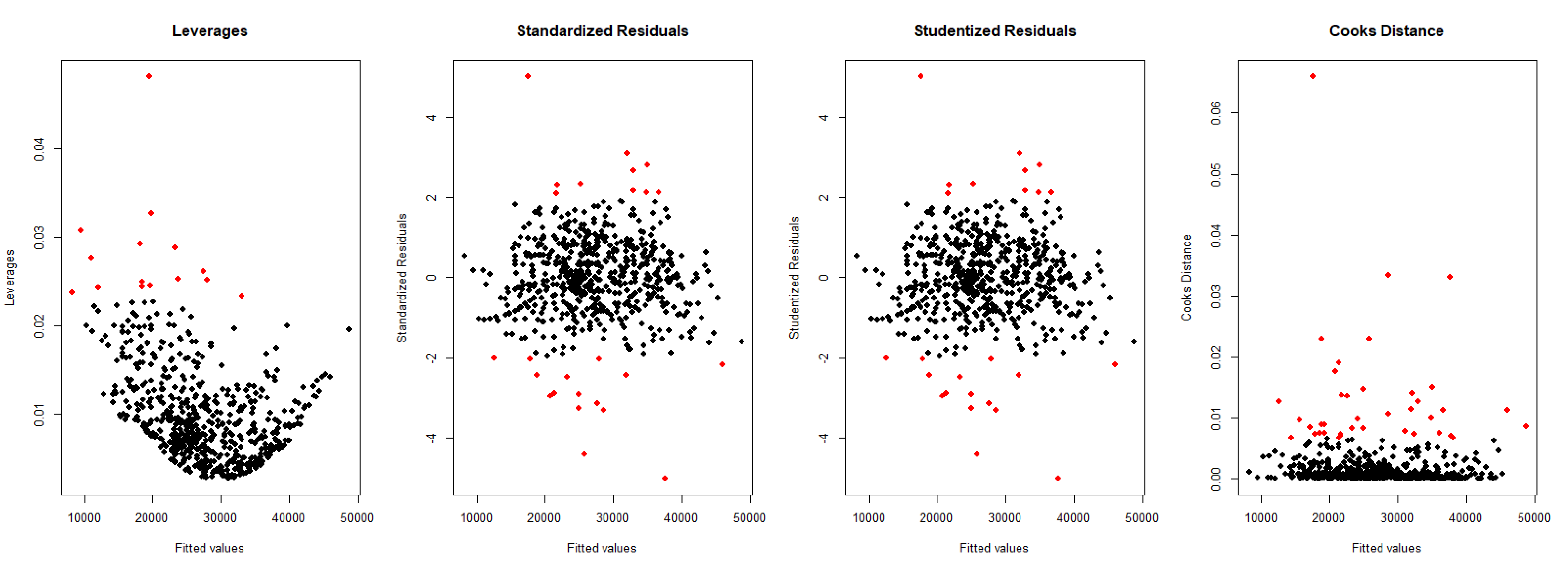
- $R^2_{\text{adj}} = 0.7794$ (it remains more or less the same)
- All remaing covariates are highly significant

Normality test of the residuals

By performing the Shapiro test on the residuals of the linear model, a p-value of 8.899e-09 is obtained.



Removal of critical datapoints



The standardized residuals and the studentized residuals coincide

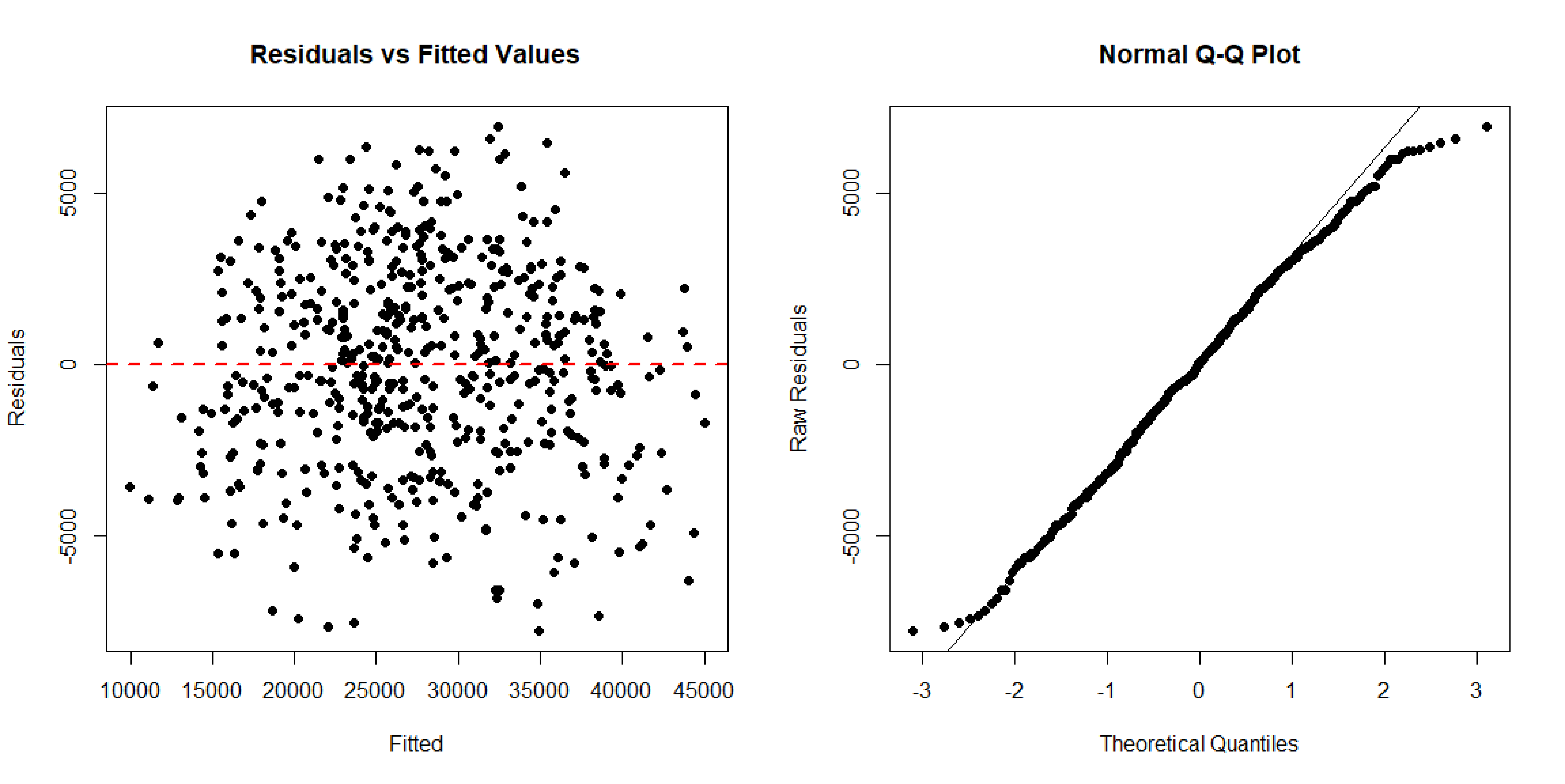
Removal of critical datapoints

Criteria	AIC	R^2_{adj}
Whole dataset	11614.97	0.779
Leverages	11333.14	0.772
Standardized residuals	10923.21	0.840
Cook's distance	10611.24	0.842
Leverage points + standardized residuals	10643.85	0.835
Standardized residuals + Cook's distance	10472.59	0.852
Leverages + Cook's distance	10193.02	0.836
Leverages + standardized residuals + Cook's distance	10213.75	0.847

The best model turns out to be the one obtained by removing the leverage points and the critical points according to the Cook's distance

New normality test of the residuals

By removing the leverage points and the critical points according to the Cook's distance, we re-train the model and obtain a p-value for the Shapiro test of 0.1387. Then, we do not reject the normality hypotesis.



Model interpretation

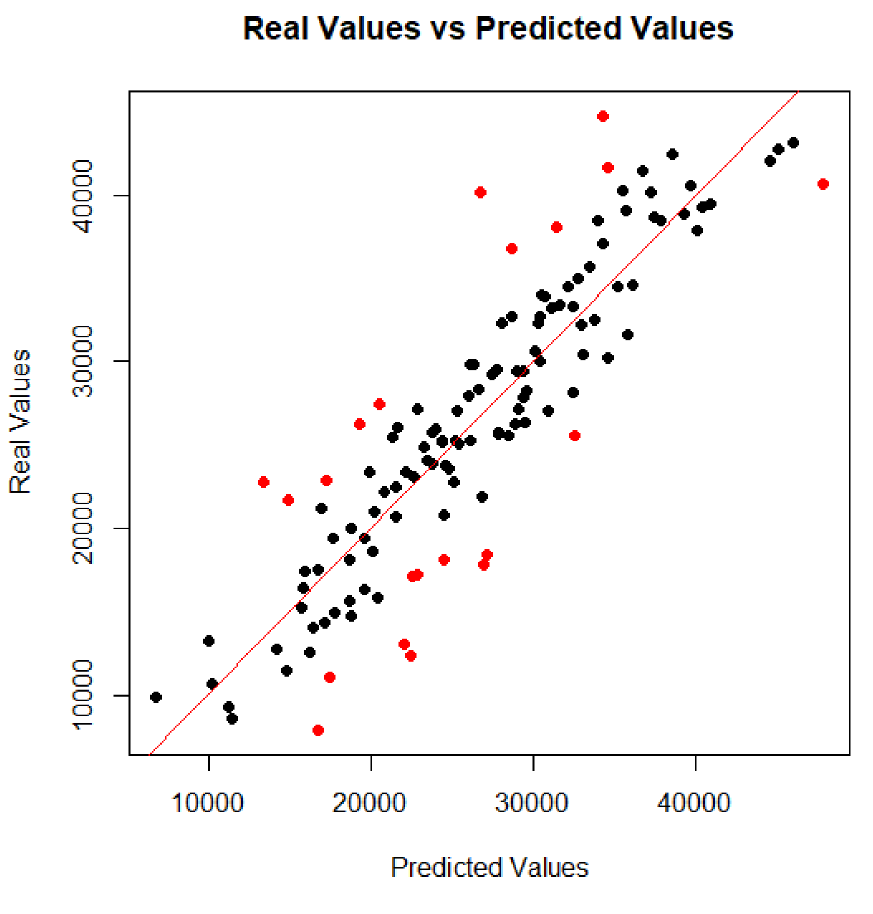
	Intercept	temp	humidity	wind_speed	is_restday	weatherrain
Coefficients β	43241	916.2	-275.5	-325.7	-5801.6	-4995.1
Coefficients β after normalization	30503	21713	-14269	-10160.4	-5801.6	-4995.1

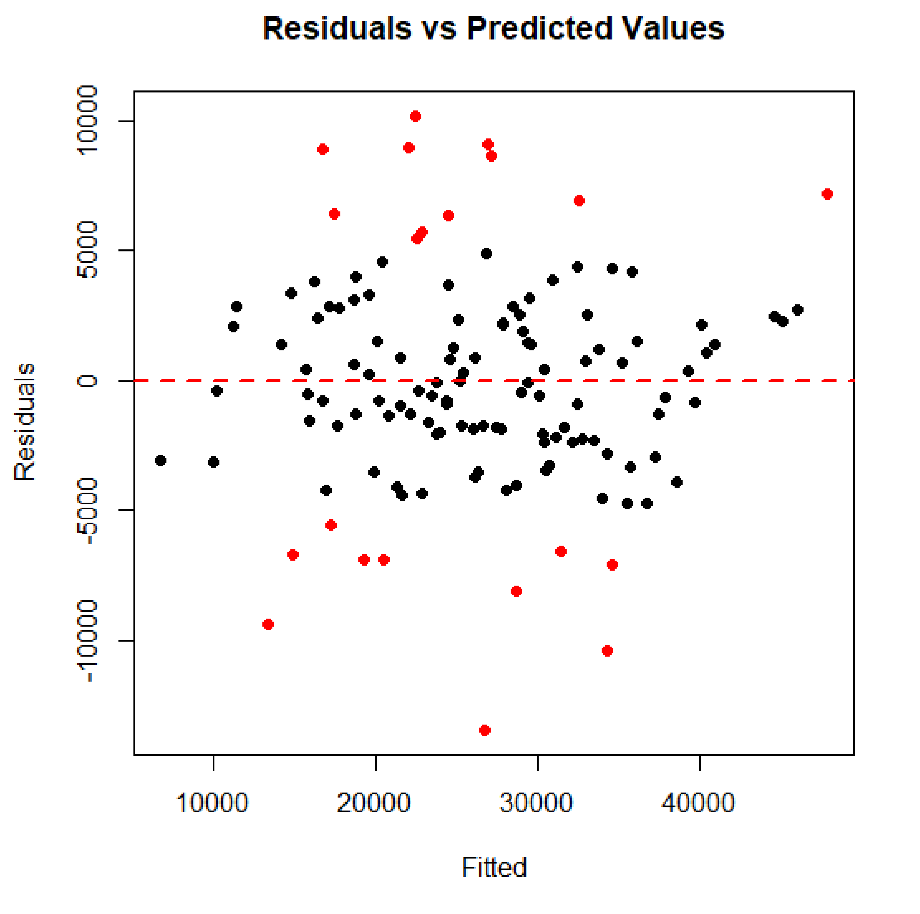
Covariate's normalization formula: $\frac{x - \min x}{\max x - \min x}$

Assessment of the model performance

To asses the model performance, we validated it on the test-set we created at the beginning.

The red dots are the datapoints that fall outside the 90% confidence interval.



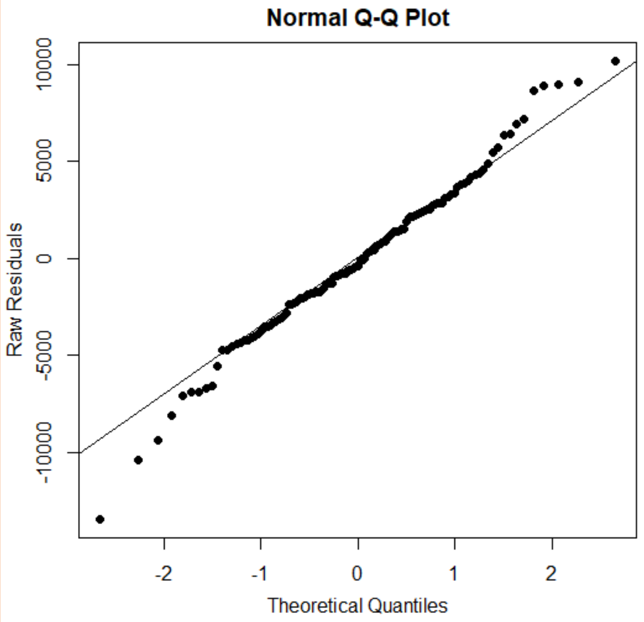


$R^2_{testset} = 0.8027$

$\rho_{real,pred} = 0.896$

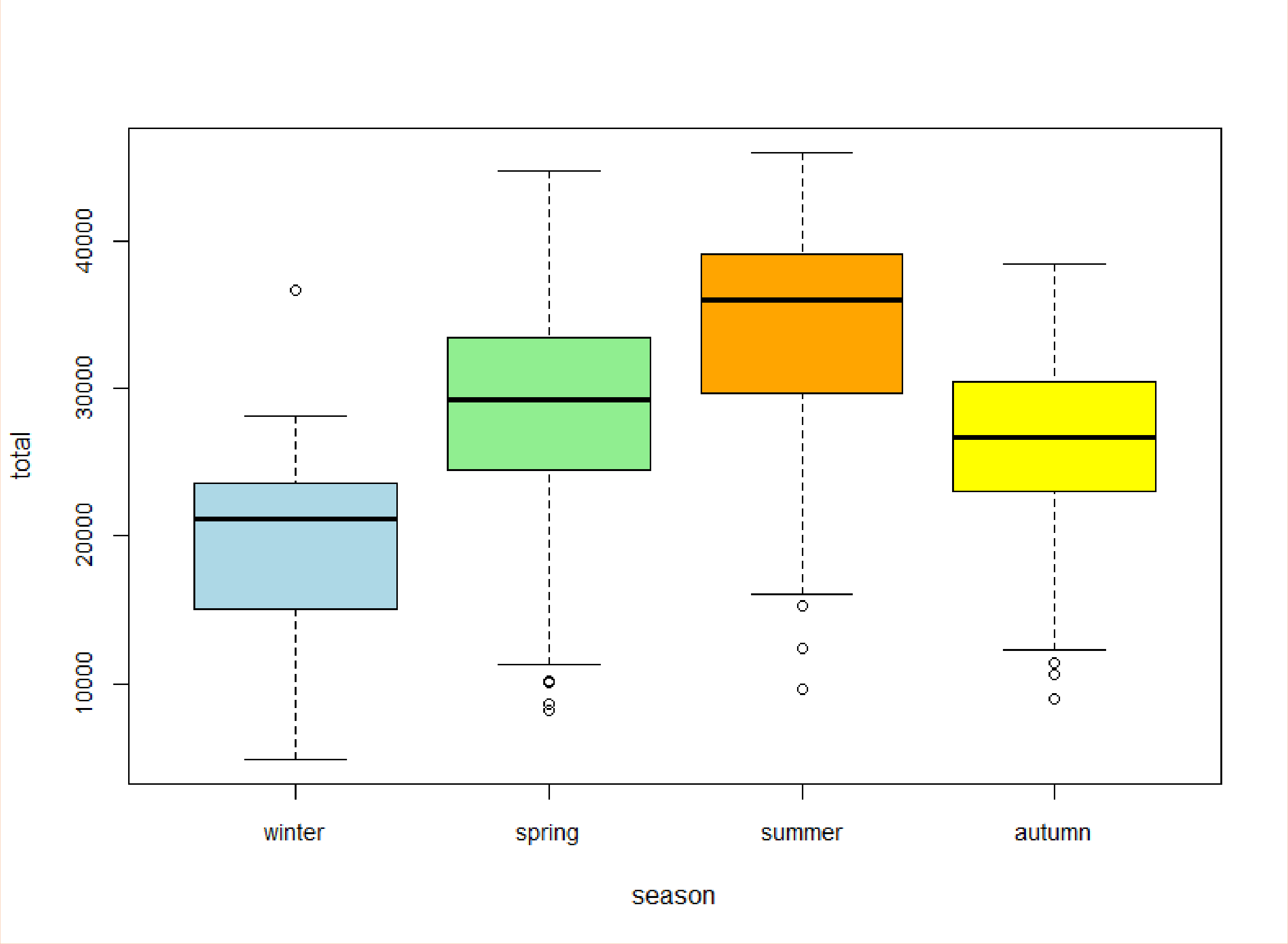
Percentage of points in the CI: 84%

P-value for the Shapiro test: 0.3371

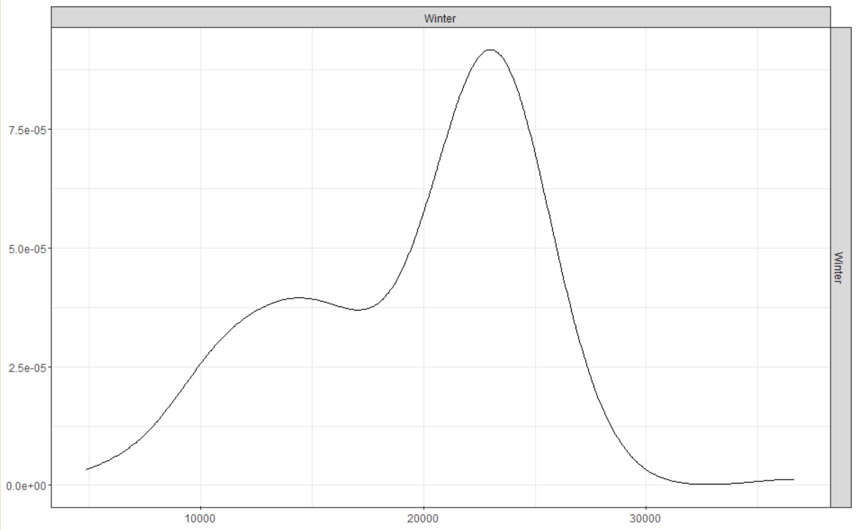


ANOVA

By observing the boxplots divided by season, we wonder if there is statistical evidence to claim that there is a difference in the average number of bikes rented in each season. We work on the entire dataset.



We limit our analysis to the Spring, Summer, and Autumn seasons. The average for Winter is visibly lower compared to the others. Moreover, the distribution of total for winter is not suitable for this analysis.

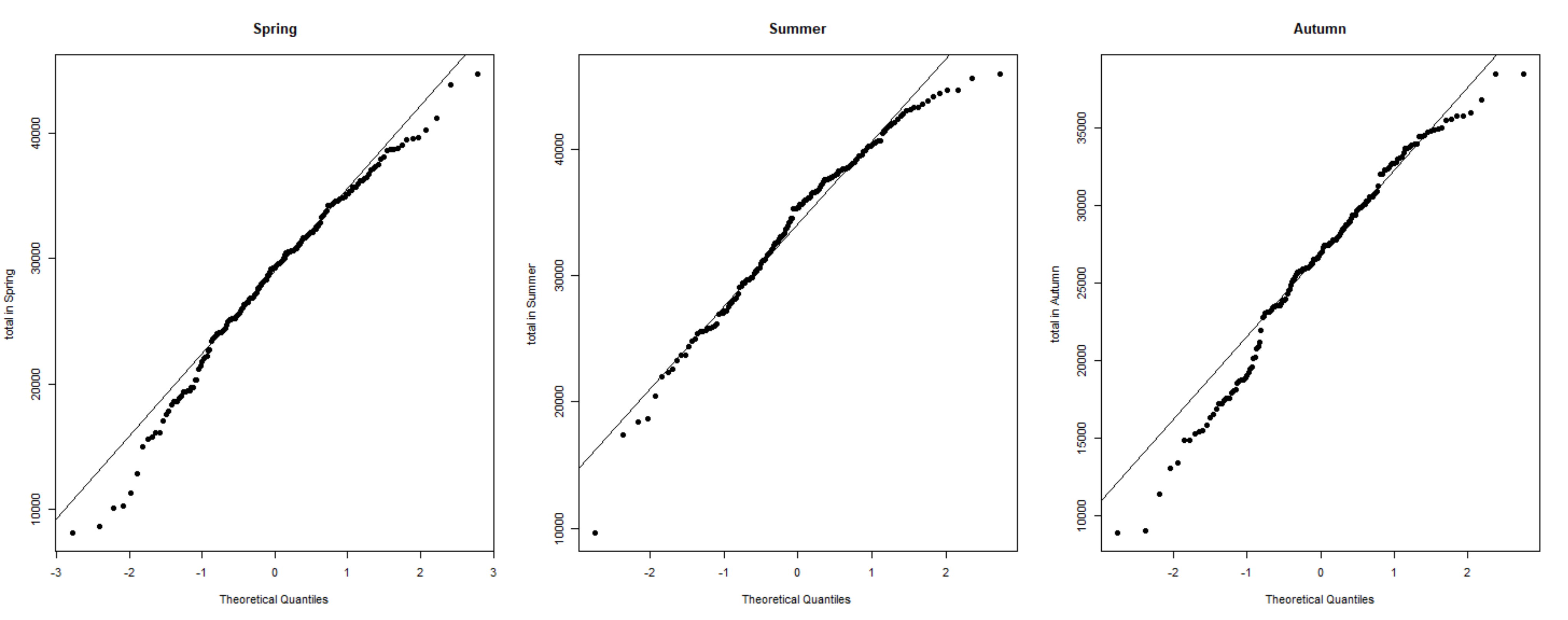


Testing of the hypoteses

By performing the Shapiro test on the total variable divided by seasons, we obtain

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

and, therefore, we reject the hypothesis of normality.



Box-Cox transformation

We try applying a Box-Cox transformation to achieve normality of total in the different seasons.

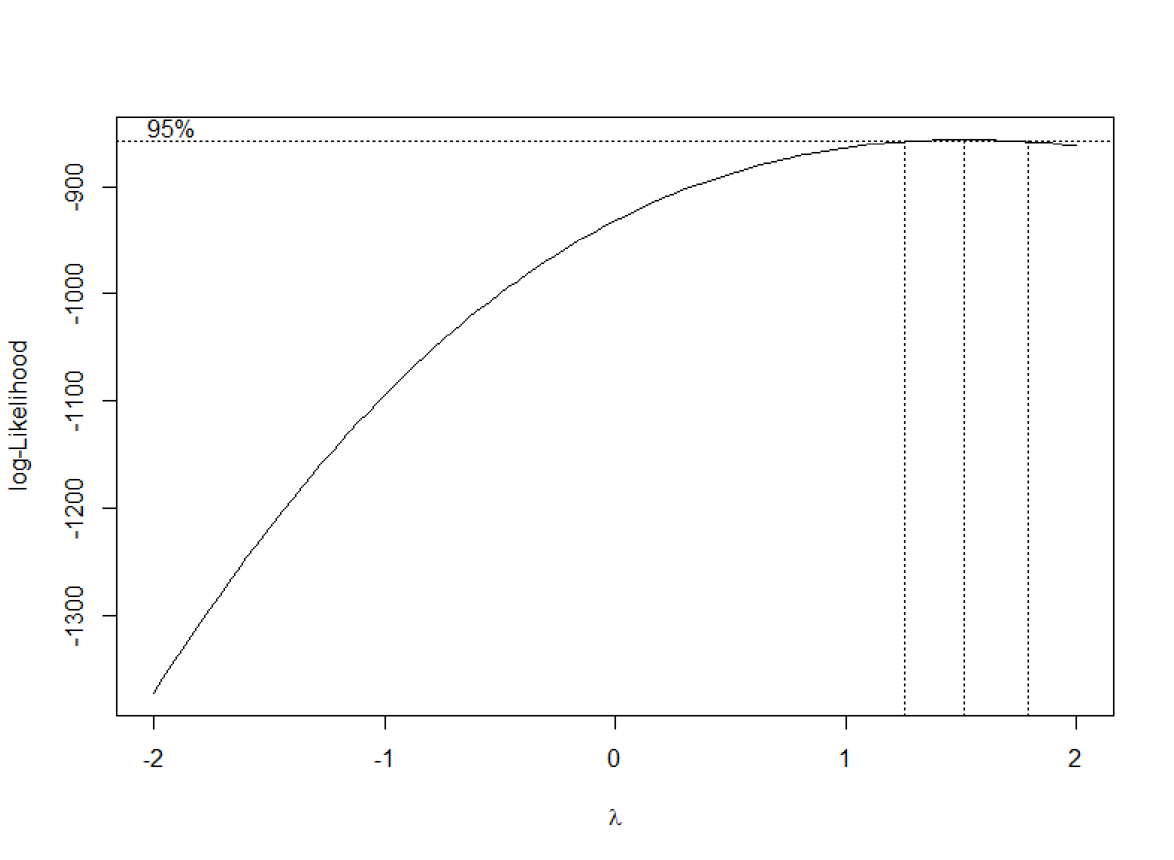
We obtain $\lambda = 1.52$

By performing the Shapiro test after the transformation, we obtain

```
> 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
```

Moreover, by performing the Bartlett test and the Levene test, we obtain

```
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

Given that the assumptions of normality and homoscedasticity are met, we can perform the ANOVA test on the groups

```
> 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
```

We reject the hypothesis that the means of the 3 groups are equal because the p-value is very low.

Finally, we perform t-tests to verify that there is a difference in the average number of bikes rented in each season.

```
> 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
```