```
setwd("/Users/lalitsachan/Desktop/March onwards/CBAP with R/Data/")
cycle_data=read.csv("Cycle_Shared.csv")
library(lubridate)
cycle_data$date=parse_date_time(cycle_data$dteday, "ymd")
cycle_data$day=day(cycle_data$date)
library(dplyr)
glimpse(cycle_data)
```

```
## Observations: 731
## Variables: 18
## $ instant    (int) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ...
## $ dteday     (fctr) 2011-01-01, 2011-01-02, 2011-01-03, 2011-01-04, 20...
## $ season     (int) 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ yr         (int) 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
## $ mnth       (int) 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ...
## $ holiday    (int) 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, ...
## $ weekday    (int) 6, 0, 1, 2, 3, 4, 5, 6, 0, 1, 2, 3, 4, 5, 6, 0, 1, ...
## $ workingday (int) 0, 0, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 0, 0, 0, ...
## $ weathersit (int) 2, 2, 1, 1, 1, 1, 2, 2, 1, 1, 2, 1, 1, 1, 2, 1, 2, ...
## $ temp       (dbl) 0.3441670, 0.3634780, 0.1963640, 0.2000000, 0.22695...
## $ atemp      (dbl) 0.3636250, 0.3537390, 0.1894050, 0.2121220, 0.22927...
## $ hum        (dbl) 0.805833, 0.696087, 0.437273, 0.590435, 0.436957, 0...
## $ windspeed  (dbl) 0.1604460, 0.2485390, 0.2483090, 0.1602960, 0.18690...
## $ casual     (int) 331, 131, 120, 108, 82, 88, 148, 68, 54, 41, 43, 25...
## $ registered (int) 654, 670, 1229, 1454, 1518, 1518, 1362, 891, 768, 1...
## $ cnt        (int) 985, 801, 1349, 1562, 1600, 1606, 1510, 959, 822, 1...
## $ date       (time) 2011-01-01, 2011-01-02, 2011-01-03, 2011-01-04, 20...
## $ day        (int) 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, ...
```

```
set.seed(2)
s=sample(1:nrow(cycle_data),0.8*nrow(cycle_data))
cd_trainval=cycle_data[s,]
cd_test=cycle_data[-s,]

s=sample(1:nrow(cd_trainval),0.7*nrow(cd_trainval))
cd_train=cd_trainval[s,]
cd_val=cd_train[-s,]

rm(s,cd_trainval)


fit_train=lm(cnt~.-instant-dteday-casual-registered-date,data=cd_train)


library(car)
vif(fit_train)
```

```
##     season         yr       mnth    holiday    weekday workingday
##   3.774689   1.046652   3.501666   1.096023   1.018826   1.099365
## weathersit       temp      atemp        hum  windspeed        day
##   1.742420 174.246951 178.097238   1.912677   1.240209   1.041207
```

```
fit_train=lm(cnt~.-instant-dteday-casual-registered-date-atemp,data=cd_train)
vif(fit_train)
```

```
##     season         yr       mnth    holiday    weekday workingday
##   3.749883   1.042988   3.500985   1.093898   1.016073   1.098460
## weathersit       temp        hum  windspeed        day
##   1.707304   1.235633   1.865266   1.177644   1.014958
```

```
fit_train=step(fit_train)
```

```
## Start:  AIC=5565.24
## cnt ~ (instant + dteday + season + yr + mnth + holiday + weekday +
##     workingday + weathersit + temp + atemp + hum + windspeed +
##     casual + registered + date + day) - instant - dteday - casual -
##     registered - date - atemp
##
##              Df Sum of Sq       RSS    AIC
## - day         1    237484 323097519 5563.5
## - mnth        1    920591 323780626 5564.4
## <none>                    322860035 5565.2
## - holiday     1   1723621 324583655 5565.4
## - hum         1   3161488 326021523 5567.2
## - weekday     1   3794998 326655033 5568.0
## - workingday  1   4120058 326980093 5568.4
## - windspeed   1  27318145 350178180 5596.4
## - weathersit  1  33343763 356203797 5603.3
## - season      1  33417603 356277637 5603.4
## - temp        1 279374561 602234596 5817.6
## - yr          1 433524472 756384506 5910.6
##
## Step:  AIC=5563.54
## cnt ~ season + yr + mnth + holiday + weekday + workingday + weathersit +
##     temp + hum + windspeed
##
##              Df Sum of Sq       RSS    AIC
## - mnth        1    923888 324021407 5562.7
## <none>                    323097519 5563.5
## - holiday     1   1680822 324778342 5563.7
## - hum         1   3302492 326400011 5565.7
## - weekday     1   3757323 326854842 5566.3
## - workingday  1   4112148 327209668 5566.7
## - windspeed   1  27371245 350468764 5594.7
## - weathersit  1  33114065 356211584 5601.3
## - season      1  33401937 356499456 5601.7
## - temp        1 279240909 602338428 5815.7
## - yr          1 436896648 759994167 5910.5
##
## Step:  AIC=5562.7
## cnt ~ season + yr + holiday + weekday + workingday + weathersit +
##     temp + hum + windspeed
##
##              Df Sum of Sq       RSS    AIC
```

```
## <none>                    324021407 5562.7
## - holiday     1   2013926 326035333 5563.2
## - weekday     1   3529030 327550437 5565.1
## - hum         1   3713029 327734436 5565.4
## - workingday  1   4285509 328306916 5566.1
## - windspeed   1  28069471 352090878 5594.6
## - weathersit  1  32692272 356713679 5599.9
## - season      1  76505555 400526963 5647.2
## - temp        1 285074533 609095940 5818.2
## - yr          1 437577172 761598579 5909.4
```

```
summary(fit_train)
```

```
##
## Call:
## lm(formula = cnt ~ season + yr + holiday + weekday + workingday +
##     weathersit + temp + hum + windspeed, data = cd_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3965.6  -461.1    23.0   550.6  3068.3
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1795.63     309.74   5.797 1.37e-08 ***
## season        426.21      43.97   9.694  < 2e-16 ***
## yr           2110.31      91.03  23.184  < 2e-16 ***
## holiday      -415.23     264.00  -1.573   0.1166
## weekday        46.02      22.10   2.082   0.0380 *
## workingday    226.44      98.69   2.294   0.0223 *
## weathersit   -662.12     104.49  -6.337 6.35e-10 ***
## temp         4971.96     265.70  18.713  < 2e-16 ***
## hum          -877.67     410.97  -2.136   0.0333 *
## windspeed   -3697.87     629.77  -5.872 9.10e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 902.3 on 398 degrees of freedom
## Multiple R-squared:    0.8,  Adjusted R-squared:  0.7955
## F-statistic: 176.9 on 9 and 398 DF,  p-value: < 2.2e-16
```

```
#removing variable holiday from the model
formula(fit_train)
```

```
## cnt ~ season + yr + holiday + weekday + workingday + weathersit +
##     temp + hum + windspeed
```

```
fit_train=lm(cnt ~ season + yr  + weekday + workingday + weathersit +
    temp + hum + windspeed,data=cd_train)
summary(fit_train)
```

```
##
```

```
## Call:
## lm(formula = cnt ~ season + yr + weekday + workingday + weathersit +
##     temp + hum + windspeed, data = cd_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3959.9  -466.4    39.4   528.3  3095.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1757.19     309.34   5.680 2.59e-08 ***
## season        424.26      44.03   9.636  < 2e-16 ***
## yr           2113.18      91.18  23.177  < 2e-16 ***
## weekday        48.64      22.08   2.203  0.02818 *
## workingday    265.12      95.76   2.769  0.00589 **
## weathersit   -659.47     104.67  -6.301 7.84e-10 ***
## temp         4973.29     266.19  18.683  < 2e-16 ***
## hum          -893.99     411.60  -2.172  0.03044 *
## windspeed   -3691.88     630.92  -5.852 1.02e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 904 on 399 degrees of freedom
## Multiple R-squared:  0.7988, Adjusted R-squared:  0.7947
## F-statistic:   198 on 8 and 399 DF,  p-value: < 2.2e-16
```

```
# lets check consistency of these variables on the validation data

fit_val=lm(cnt ~.-instant-dteday-casual-registered-date,data=cd_val)
vif(fit_val)
```

```
##     season         yr       mnth    holiday    weekday workingday
##   5.482866   1.121207   4.955398   1.217113   1.100220   1.247158
## weathersit       temp      atemp        hum  windspeed        day
##   2.437789 160.452822 163.219708   2.737067   1.360679   1.056929
```

```
fit_val=lm(cnt ~.-instant-dteday-casual-registered-date-atemp,data=cd_val)
vif(fit_val)
```

```
##     season         yr       mnth    holiday    weekday workingday
##   5.472686   1.099030   4.955392   1.216411   1.097467   1.233948
## weathersit       temp        hum  windspeed        day
##   2.355899   1.356788   2.611746   1.358753   1.046020
```

```
fit_val=step(fit_val)
```

```
## Start:  AIC=1783.36
## cnt ~ (instant + dteday + season + yr + mnth + holiday + weekday +
##     workingday + weathersit + temp + atemp + hum + windspeed +
##     casual + registered + date + day) - instant - dteday - casual -
##     registered - date - atemp
##
```

```
##              Df Sum of Sq       RSS    AIC
## - holiday     1     17405 108082254 1781.4
## - mnth        1    115026 108179875 1781.5
## - day         1   1171867 109236716 1782.8
## - weekday     1   1313091 109377939 1782.9
## - season      1   1369320 109434168 1783.0
## <none>                    108064849 1783.4
## - hum         1   2821648 110886497 1784.7
## - workingday  1   3880545 111945394 1785.9
## - weathersit  1   5025608 113090456 1787.2
## - windspeed   1  12089660 120154509 1795.0
## - temp        1  96999847 205064696 1864.0
## - yr          1 147284509 255349358 1892.3
##
## Step:  AIC=1781.38
## cnt ~ season + yr + mnth + weekday + workingday + weathersit +
##     temp + hum + windspeed + day
##
##              Df Sum of Sq       RSS    AIC
## - mnth        1    143212 108225466 1779.5
## - day         1   1166548 109248802 1780.8
## - weekday     1   1317575 109399829 1780.9
## - season      1   1379191 109461445 1781.0
## <none>                    108082254 1781.4
## - hum         1   2855645 110937899 1782.7
## - workingday  1   4087415 112169669 1784.2
## - weathersit  1   5015249 113097503 1785.2
## - windspeed   1  12220283 120302537 1793.2
## - temp        1 102463233 210545487 1865.4
## - yr          1 150651663 258733916 1892.0
##
## Step:  AIC=1779.55
## cnt ~ season + yr + weekday + workingday + weathersit + temp +
##     hum + windspeed + day
##
##              Df Sum of Sq       RSS    AIC
## - day         1   1253354 109478820 1779.0
## - weekday     1   1311720 109537186 1779.1
## <none>                    108225466 1779.5
## - hum         1   2755735 110981201 1780.8
## - workingday  1   4020946 112246412 1782.2
## - weathersit  1   5074904 113300370 1783.5
## - season      1   8614425 116839891 1787.4
## - windspeed   1  12086408 120311874 1791.2
## - temp        1 102333893 210559359 1863.4
## - yr          1 151233004 259458470 1890.3
##
## Step:  AIC=1779.03
## cnt ~ season + yr + weekday + workingday + weathersit + temp +
##     hum + windspeed
##
##              Df Sum of Sq       RSS    AIC
## - weekday     1   1306658 110785478 1778.6
## <none>                    109478820 1779.0
```

```
## - hum         1   2568791 112047611 1780.0
## - workingday  1   4350216 113829036 1782.1
## - weathersit  1   5341095 114819915 1783.2
## - season      1   7894947 117373767 1786.0
## - windspeed   1  12301541 121780361 1790.8
## - temp        1 101322830 210801650 1861.5
## - yr          1 151215409 260694230 1889.0
##
## Step:  AIC=1778.56
## cnt ~ season + yr + workingday + weathersit + temp + hum + windspeed
##
##              Df Sum of Sq       RSS    AIC
## <none>                     110785478 1778.6
## - hum         1   2349887 113135365 1779.3
## - weathersit  1   5166366 115951844 1782.4
## - workingday  1   5499356 116284834 1782.8
## - season      1   7861041 118646519 1785.4
## - windspeed   1  11396698 122182176 1789.2
## - temp        1 101420113 212205591 1860.4
## - yr          1 150590340 261375818 1887.3
```

```
summary(fit_val)
```

```
##
## Call:
## lm(formula = cnt ~ season + yr + workingday + weathersit + temp +
##     hum + windspeed, data = cd_val)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3864.3  -543.4   104.6   572.3  2609.9
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2270.50     584.21   3.886 0.000167 ***
## season        262.44      89.56   2.930 0.004051 **
## yr           2236.89     174.42  12.825  < 2e-16 ***
## workingday    452.04     184.44   2.451 0.015684 *
## weathersit   -559.95     235.72  -2.375 0.019099 *
## temp         5424.69     515.42  10.525  < 2e-16 ***
## hum         -1476.44     921.59  -1.602 0.111753
## windspeed   -4273.27    1211.21  -3.528 0.000592 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 956.9 on 121 degrees of freedom
## Multiple R-squared:  0.8066, Adjusted R-squared:  0.7955
## F-statistic: 72.11 on 7 and 121 DF,  p-value: < 2.2e-16
```

```
formula(fit_val)
```

```
## cnt ~ season + yr + workingday + weathersit + temp + hum + windspeed
```

```r
#removing variable hum for high prob value

fit_val=lm(cnt ~ season + yr + workingday + weathersit + temp + windspeed,data=cd_val)

summary(fit_val)
```

```
##
## Call:
## lm(formula = cnt ~ season + yr + workingday + weathersit + temp +
##     windspeed, data = cd_val)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3920.7  -552.4   114.0   610.5  2619.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1640.23     434.65   3.774  0.00025 ***
## season        243.00      89.31   2.721  0.00746 **
## yr           2288.43     172.53  13.264  < 2e-16 ***
## workingday    461.25     185.53   2.486  0.01427 *
## weathersit   -835.50     162.22  -5.150 1.01e-06 ***
## temp         5352.98     516.76  10.359  < 2e-16 ***
## windspeed   -3539.96    1128.56  -3.137  0.00214 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 963 on 122 degrees of freedom
## Multiple R-squared:  0.8025, Adjusted R-squared:  0.7928
## F-statistic: 82.64 on 6 and 122 DF,  p-value: < 2.2e-16
```

```r
summary(fit_train)
```

```
##
## Call:
## lm(formula = cnt ~ season + yr + weekday + workingday + weathersit +
##     temp + hum + windspeed, data = cd_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3959.9  -466.4    39.4   528.3  3095.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1757.19     309.34   5.680 2.59e-08 ***
## season        424.26      44.03   9.636  < 2e-16 ***
## yr           2113.18      91.18  23.177  < 2e-16 ***
## weekday        48.64      22.08   2.203  0.02818 *
## workingday    265.12      95.76   2.769  0.00589 **
## weathersit   -659.47     104.67  -6.301 7.84e-10 ***
## temp         4973.29     266.19  18.683  < 2e-16 ***
## hum          -893.99     411.60  -2.172  0.03044 *
## windspeed   -3691.88     630.92  -5.852 1.02e-08 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 904 on 399 degrees of freedom
## Multiple R-squared:  0.7988, Adjusted R-squared:  0.7947
## F-statistic:   198 on 8 and 399 DF,  p-value: < 2.2e-16
```

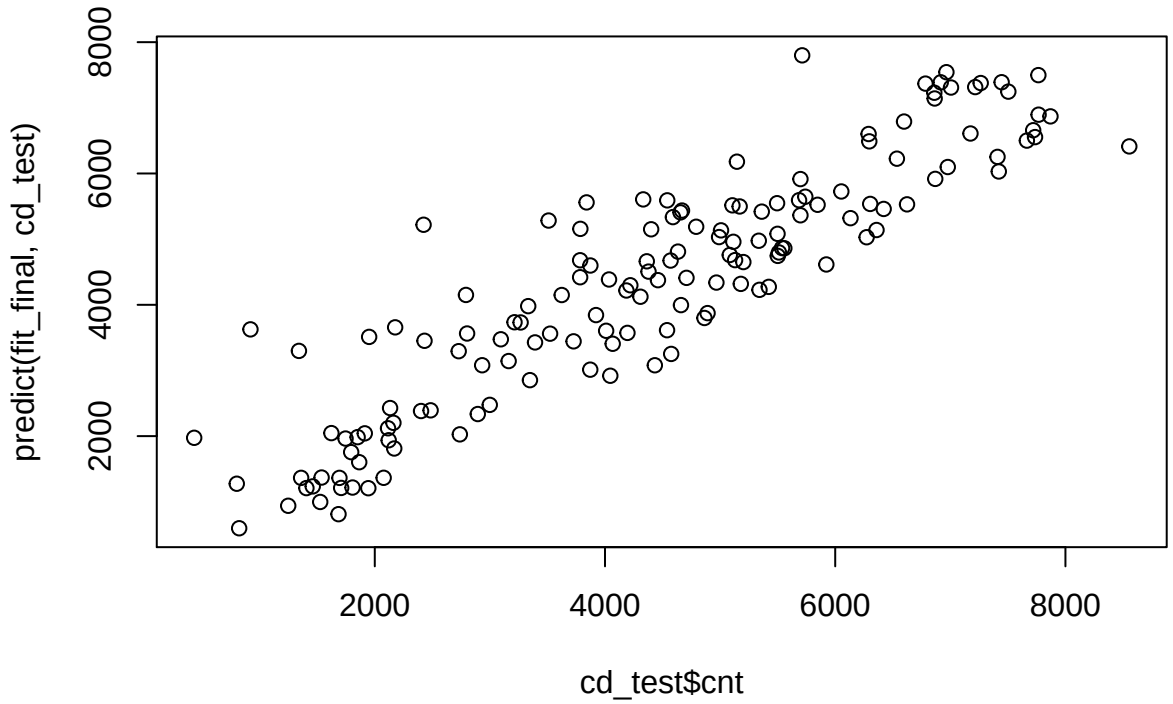
```
#picking consistent variable from the comparison and building model with those on train data
fit_final=lm(cnt ~ season + yr + workingday + weathersit + temp +  windspeed,data=cd_train)

summary(fit_final)
```

```
##
## Call:
## lm(formula = cnt ~ season + yr + workingday + weathersit + temp +
##     windspeed, data = cd_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4143.9  -481.8    57.6   584.1  3219.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1500.23     246.40   6.089 2.67e-09 ***
## season        415.79      44.26   9.394  < 2e-16 ***
## yr           2149.86      91.02  23.620  < 2e-16 ***
## workingday    279.50      96.61   2.893  0.00402 **
## weathersit   -796.59      83.17  -9.578  < 2e-16 ***
## temp         4882.76     265.15  18.415  < 2e-16 ***
## windspeed   -3309.60     612.43  -5.404 1.12e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 913.1 on 401 degrees of freedom
## Multiple R-squared:  0.7936, Adjusted R-squared:  0.7906
## F-statistic:   257 on 6 and 401 DF,  p-value: < 2.2e-16
```

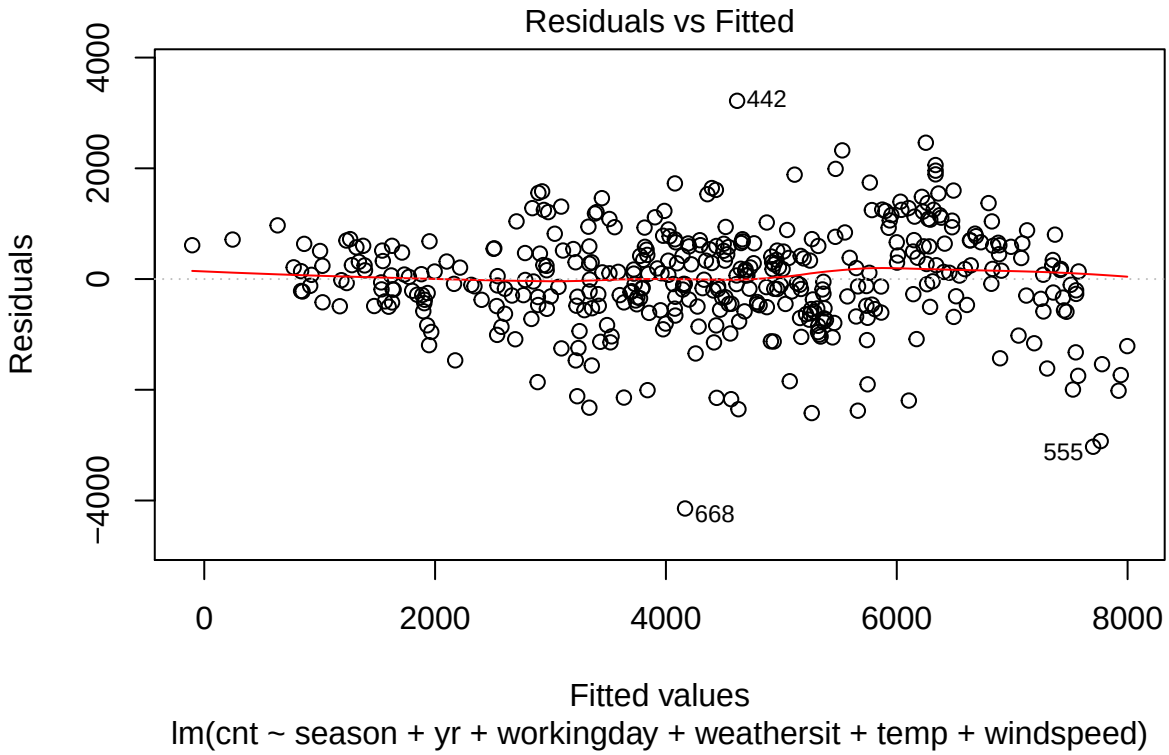
```
#Performance on test data
rmse=sqrt(mean((predict(fit_final,cd_test)-cd_test$cnt)**2))
rmse
```

```
## [1] 823.7172
```

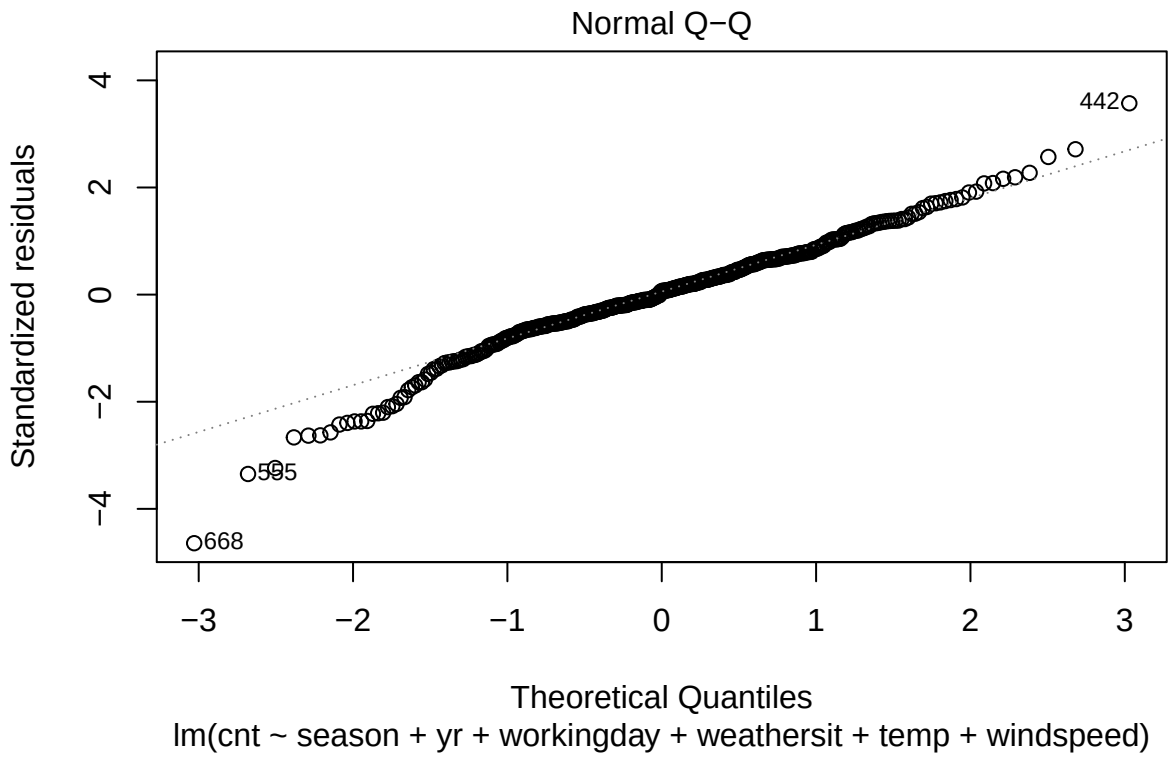
```
# visual agreement
plot(cd_test$cnt,predict(fit_final,cd_test))
```



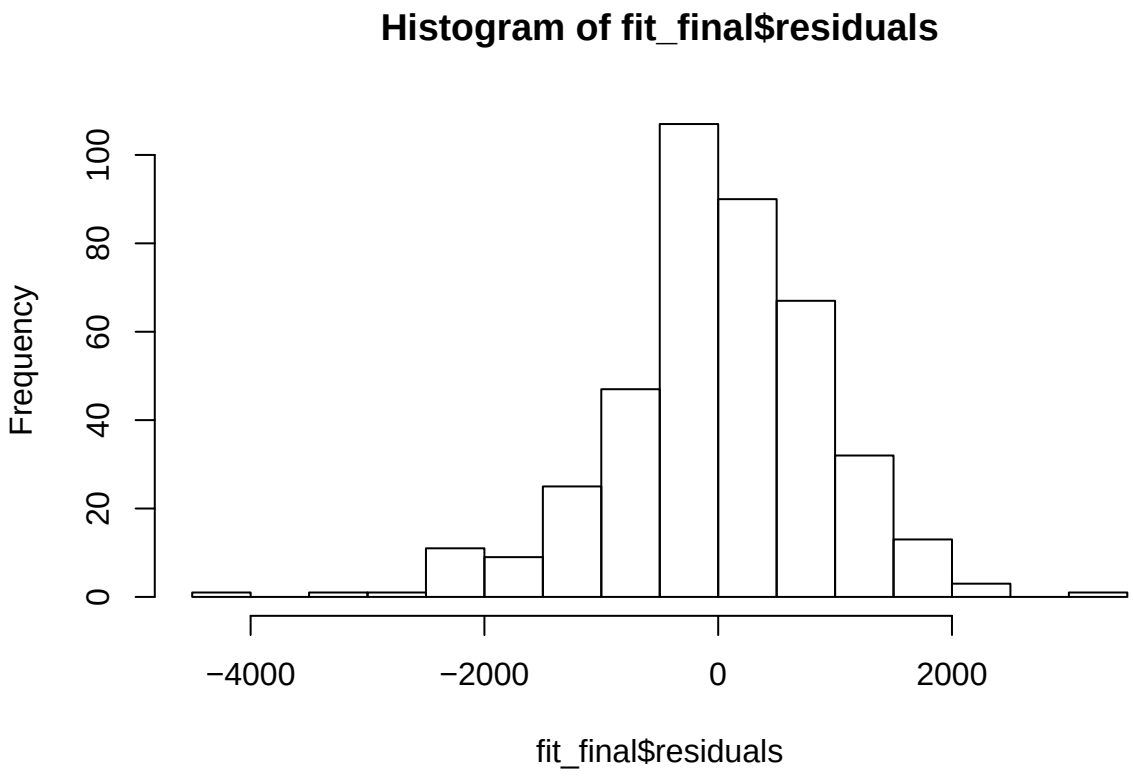

```
# Fit diagnostics

# Error randomness

plot(fit_final,which=1)
```



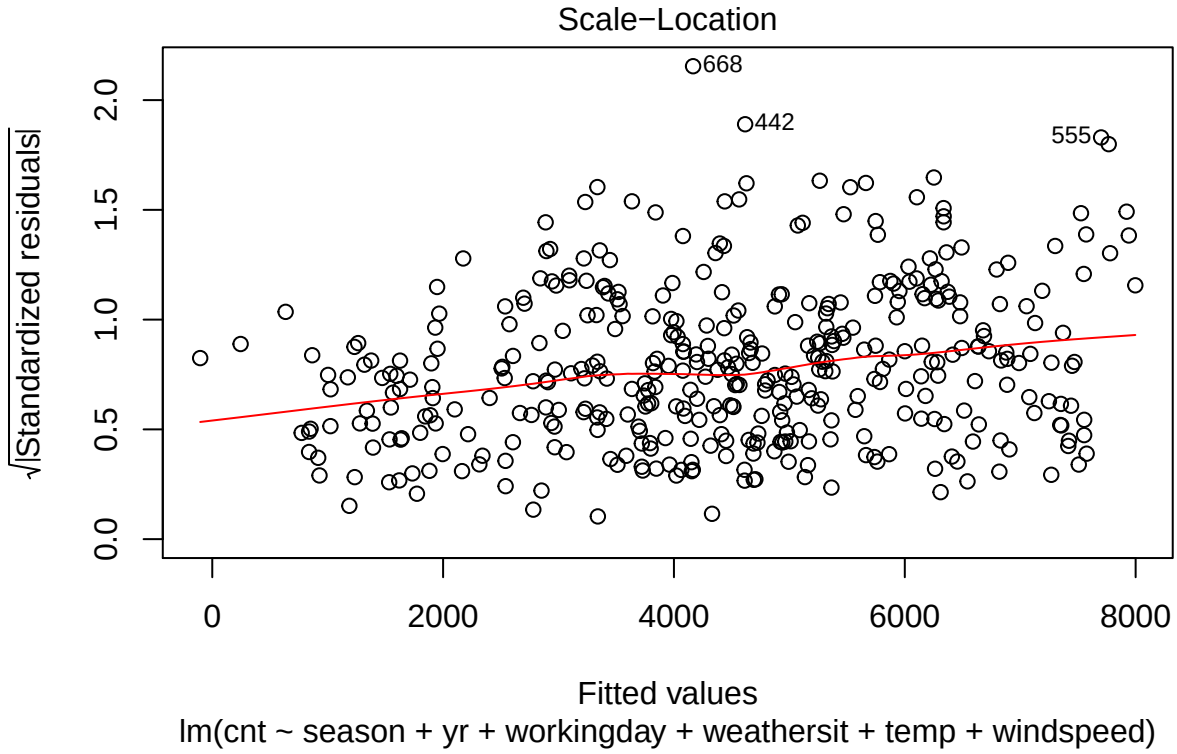
```r
# There doesnt seem to be any pattern , we need not worry about our model definition

# Error Normality

plot(fit_final,which=2)
```



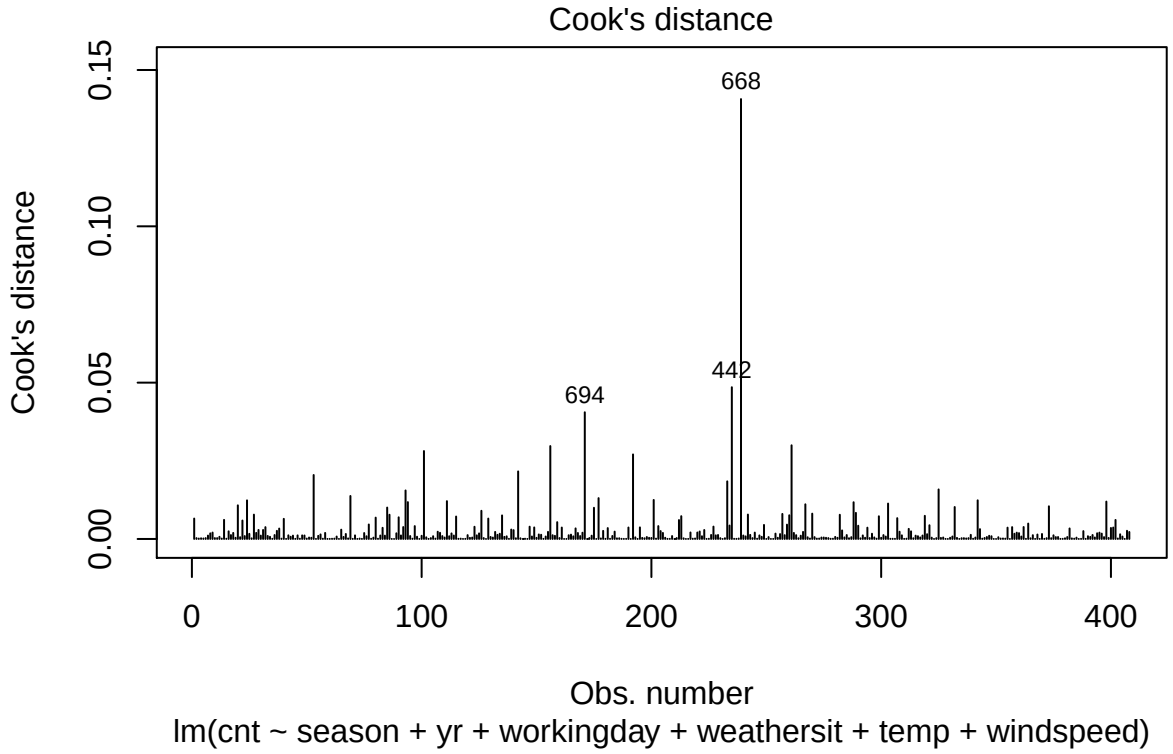

```r
hist(fit_final$residuals,breaks = 20)
```




```
# Error variance
plot(fit_final,which=3)
```



```
# Outliers detection : None found , cook's distance < 1 for all obs

plot(fit_final,which=4)
```



Discuss your views/doubts on fit diagnostic plots on QA forum.