LAB2RMD

```
# Duomenys
# https://www.kaggle.com/datasets/brajeshmohapatra/bike-count-prediction-data-set?select=train.csv

library(tidyverse)
library(ggplot2)
library(reshape2)
library(AER)
library(MASS)
```

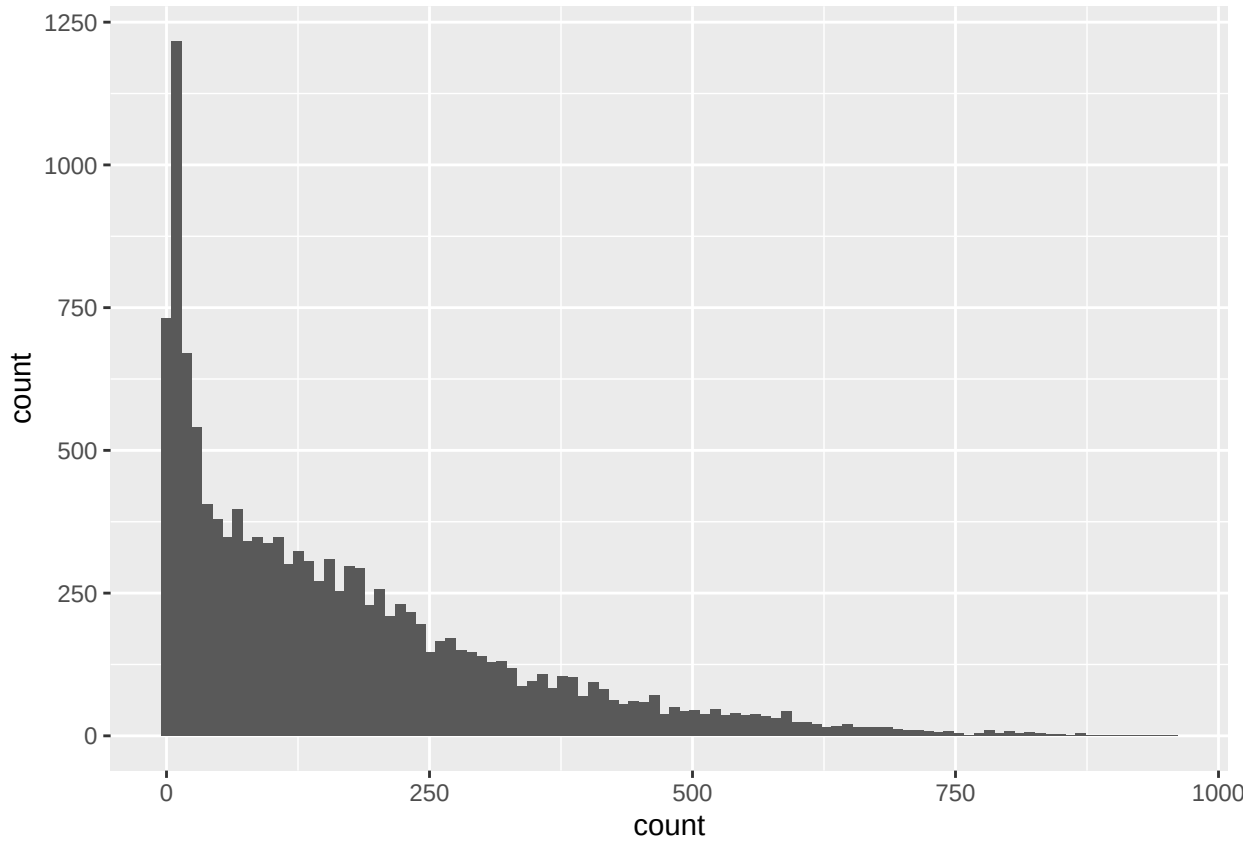
```
d <- read.csv("train.csv")
d <- dplyr::select(d, -c(datetime, casual, registered))

head(d)
```

```
##   season holiday workingday weather temp  atemp humidity windspeed count
## 1      1       0          0       1 9.84 14.395       81    0.0000    16
## 2      1       0          0       1 9.02 13.635       80    0.0000    40
## 3      1       0          0       1 9.02 13.635       80    0.0000    32
## 4      1       0          0       1 9.84 14.395       75    0.0000    13
## 5      1       0          0       1 9.84 14.395       75    0.0000     1
## 6      1       0          0       2 9.84 12.880       75    6.0032     1
```

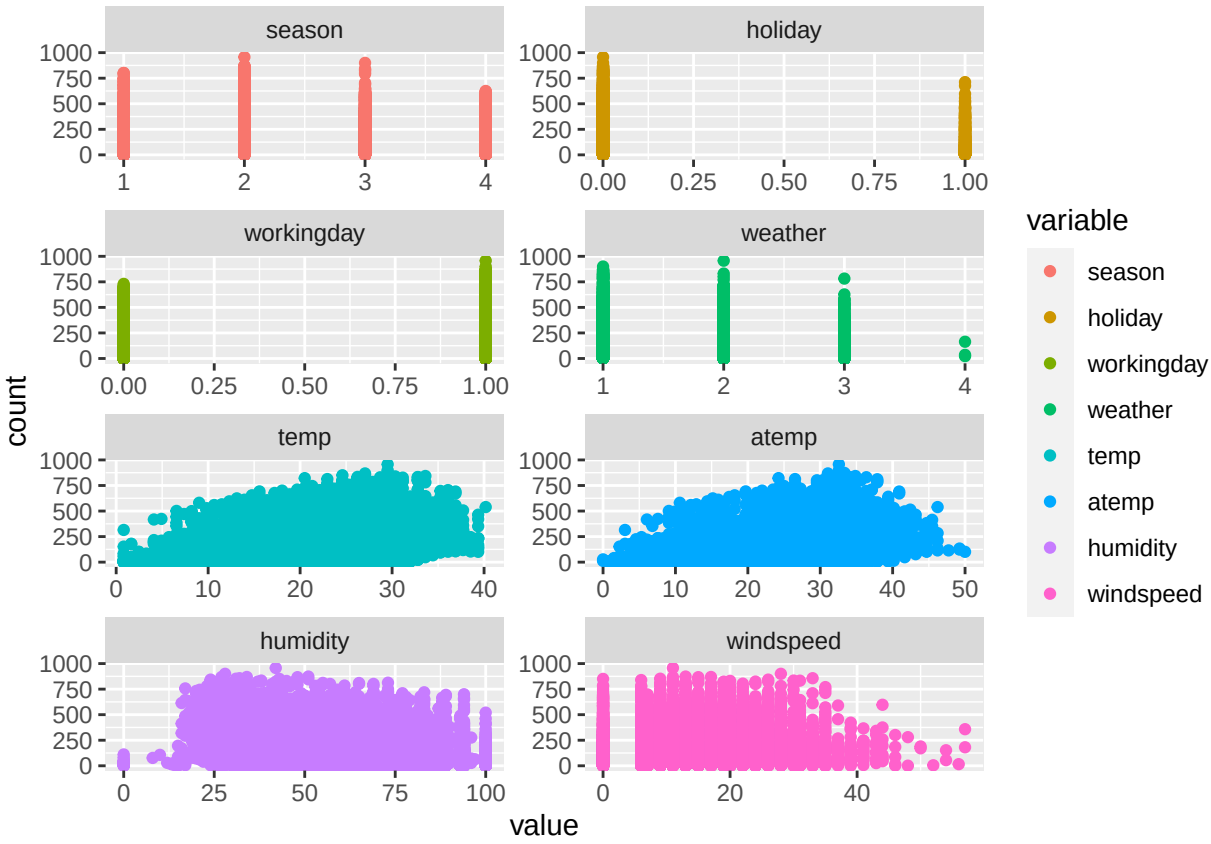
```
summary(d)
```

```
##      season         holiday         workingday       weather
##  Min.   :1.000   Min.   :0.0000   Min.   :0.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:1.000
##  Median :2.000   Median :0.0000   Median :1.000   Median :1.000
##  Mean   :2.211   Mean   :0.0275   Mean   :0.686   Mean   :1.427
##  3rd Qu.:3.000   3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:2.000
##  Max.   :4.000   Max.   :1.0000   Max.   :1.000   Max.   :4.000
##       temp           atemp          humidity       windspeed
##  Min.   : 0.82   Min.   : 0.00   Min.   :  0.00   Min.   : 0.000
##  1st Qu.:13.12   1st Qu.:15.91   1st Qu.: 47.00   1st Qu.: 7.002
##  Median :19.68   Median :23.48   Median : 62.00   Median :12.998
##  Mean   :19.73   Mean   :23.11   Mean   : 62.36   Mean   :13.142
##  3rd Qu.:26.24   3rd Qu.:30.30   3rd Qu.: 79.00   3rd Qu.:19.001
##  Max.   :40.18   Max.   :50.00   Max.   :100.00   Max.   :56.997
##      count
##  Min.   :  1.0
##  1st Qu.: 35.0
##  Median :124.0
##  Mean   :167.6
##  3rd Qu.:245.0
##  Max.   :957.0
```

```
ggplot(d, aes(x=count)) + geom_histogram(bins = 100)
```




```
ggplot(melt(d, "count"), aes(x = value, y = count, colour = variable)) +
  geom_point() +
  facet_wrap(~variable, scales = "free", nrow = 4)
```




```
### Poisson
m1 <- glm(count ~ ., family="poisson", data=d)
summary(m1)
```

```
##
## Call:
## glm(formula = count ~ ., family = "poisson", data = d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -27.345   -9.535   -2.690    4.545   41.151
##
## Coefficients:
##               Estimate Std. Error  z value Pr(>|z|)
## (Intercept)  4.782e+00  3.983e-03 1200.614   <2e-16 ***
## season       8.928e-03  7.960e-04   11.216   <2e-16 ***
## holiday     -1.543e-01  4.662e-03  -33.092   <2e-16 ***
## workingday  -1.327e-02  1.518e-03   -8.740   <2e-16 ***
## weather     -1.260e-02  1.288e-03   -9.781   <2e-16 ***
## temp        -1.618e-02  6.468e-04  -25.010   <2e-16 ***
## atemp        5.846e-02  5.958e-04   98.126   <2e-16 ***
## humidity    -1.384e-02  4.098e-05 -337.640   <2e-16 ***
## windspeed    4.761e-03  8.596e-05   55.391   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
```

```
##
##     Null deviance: 1911146  on 12979  degrees of freedom
## Residual deviance: 1387988  on 12971  degrees of freedom
## AIC: 1469276
##
## Number of Fisher Scoring iterations: 5
```

```
cat("Deviacija padalinta is laisves laipsniu: ",m1$deviance / m1$df.residual)
```

```
## Deviacija padalinta is laisves laipsniu:  107.007
```

```
cat("Turi buti tarp 0.7 ir 1.3, tad nebegalime naudoti puasono modelio")
```

```
## Turi buti tarp 0.7 ir 1.3, tad nebegalime naudoti puasono modelio
```

```
dispersiontest(m1)
```

```
##
##  Overdispersion test
##
## data:  m1
## z = 51.787, p-value < 2.2e-16
## alternative hypothesis: true dispersion is greater than 1
## sample estimates:
## dispersion
##   114.8725
```

```
### Negative Binomial
m2 <- glm.nb(count ~ ., data = d)
summary(m2)
```

```
##
## Call:
## glm.nb(formula = count ~ ., data = d, init.theta = 1.030517395,
##     link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9313  -0.9845  -0.2273   0.3475   3.5581
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.5457388  0.0494949  91.843  < 2e-16 ***
## season       0.0179194  0.0092530   1.937 0.052793 .
## holiday     -0.1578975  0.0549495  -2.874 0.004059 **
## workingday   0.0997106  0.0193834   5.144 2.69e-07 ***
## weather      0.0060803  0.0151710   0.401 0.688580
## temp        -0.0303578  0.0092194  -3.293 0.000992 ***
## atemp        0.0769779  0.0084421   9.118  < 2e-16 ***
## humidity    -0.0145218  0.0005224 -27.796  < 2e-16 ***
## windspeed    0.0043411  0.0011602   3.742 0.000183 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.0305) family taken to be 1)
##
##     Null deviance: 18362  on 12979  degrees of freedom
```

```
## Residual deviance: 14867  on 12971  degrees of freedom
## AIC: 155611
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.0305
##           Std. Err.:  0.0116
##
##  2 x log-likelihood:  -155591.0130
```

```
cat("Deviacija padalinta is laisves laipsniu: ",m2$deviance / m2$df.residual)
```

```
## Deviacija padalinta is laisves laipsniu:  1.146175
```

```
### Stepwise
m2step <- stepAIC(m2, direction = "both")
```

```
## Start:  AIC=155609
## count ~ season + holiday + workingday + weather + temp + atemp +
##     humidity + windspeed
##
##              Df    AIC
## - weather     1 155607
## <none>          155609
## - season      1 155611
## - holiday     1 155615
## - temp        1 155617
## - windspeed   1 155621
## - workingday  1 155633
## - atemp       1 155682
## - humidity    1 156325
##
## Step:  AIC=155607.2
## count ~ season + holiday + workingday + temp + atemp + humidity +
##     windspeed
##
##              Df    AIC
## <none>          155607
## + weather     1 155609
## - season      1 155609
## - holiday     1 155613
## - temp        1 155615
## - windspeed   1 155619
## - workingday  1 155631
## - atemp       1 155680
## - humidity    1 156506
```

```
summary(m2step)
```

```
##
## Call:
## glm.nb(formula = count ~ season + holiday + workingday + temp +
##     atemp + humidity + windspeed, data = d, init.theta = 1.030507956,
##     link = log)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.9261  -0.9854  -0.2276   0.3477   3.5582
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.5486115  0.0489952  92.838  < 2e-16 ***
## season       0.0177488  0.0092342   1.922 0.054597 .
## holiday     -0.1580013  0.0549497  -2.875 0.004035 **
## workingday   0.0998662  0.0193611   5.158 2.49e-07 ***
## temp        -0.0302577  0.0092158  -3.283 0.001026 **
## atemp        0.0768636  0.0084370   9.110  < 2e-16 ***
## humidity    -0.0144238  0.0004649 -31.026  < 2e-16 ***
## windspeed    0.0043896  0.0011492   3.820 0.000134 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(1.0305) family taken to be 1)
##
##     Null deviance: 18362  on 12979  degrees of freedom
## Residual deviance: 14867  on 12972  degrees of freedom
## AIC: 155609
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  1.0305
##           Std. Err.:  0.0116
##
##  2 x log-likelihood:  -155591.1600
```

```
cat("Deviacija padalinta is laisves laipsniu: ",m2step$deviance / m2step$df.residual)
```

```
## Deviacija padalinta is laisves laipsniu:  1.146088
```

```
# Anova between models
anova(m2, m2step)
```

```
## Likelihood ratio tests of Negative Binomial Models
##
## Response: count
##                                                                          Model
## 1           season + holiday + workingday + temp + atemp + humidity + windspeed
## 2 season + holiday + workingday + weather + temp + atemp + humidity + windspeed
##      theta Resid. df    2 x log-lik.   Test    df  LR stat.   Pr(Chi)
## 1 1.030508     12972       -155591.2
## 2 1.030517     12971       -155591.0 1 vs 2     1 0.1463433 0.7020546
```

```
# Koeficientai
est <- cbind(Estimate = coef(m2step), confint(m2step))
```

```
## Waiting for profiling to be done...
```

```
exp(est)
```

```
##               Estimate      2.5 %      97.5 %
## (Intercept) 94.5011015 85.4811031 104.5054522
```

```
## season       1.0179073  1.0005701   1.0356400
## holiday      0.8538486  0.7677649   0.9525380
## workingday   1.1050231  1.0636883   1.1477461
## temp         0.9701955  0.9519689   0.9887905
## atemp        1.0798948  1.0612557   1.0988513
## humidity     0.9856797  0.9847646   0.9865953
## windspeed    1.0043992  1.0020931   1.0067166
```

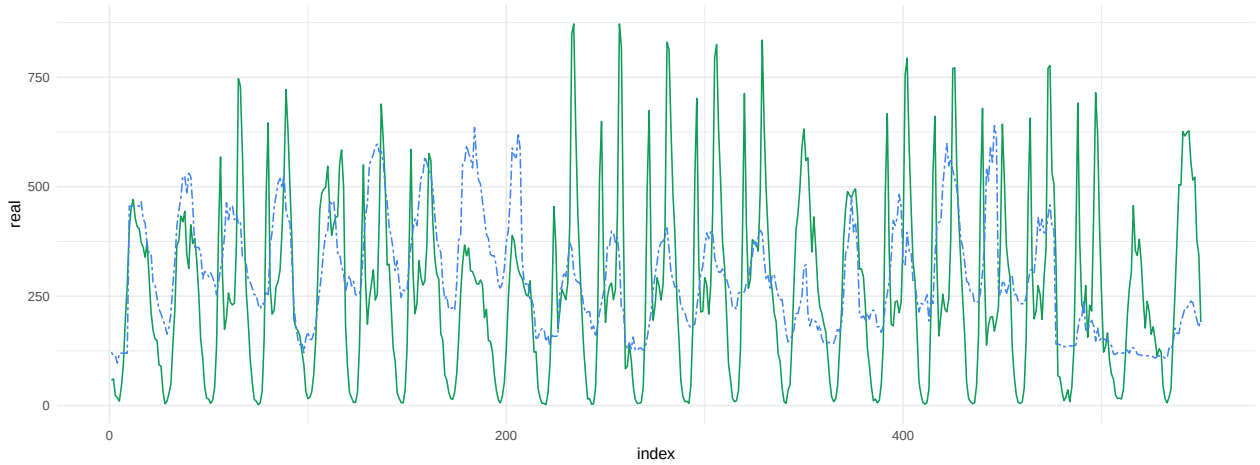
```
library(ggplot2)
theme_set(theme_minimal())
### Prediction

dopred <- function(tt, model) {
  tt$count <- tt$casual + tt$registered
  index <- tt$index
  real <- tt$count
  tt <- dplyr::select(tt, -c(datetime, casual, registered, count))
  predicted <- predict(m2step, newdata = tt, type = "response")
  tempdf <- data.frame(index, real, predicted)

  p<- ggplot(tempdf, aes(x=index)) +
    geom_line(aes(y = real), color = "#0F9D58") +
    geom_line(aes(y = predicted), color="#4285F4", linetype="twodash")
  return(p)
}


test <- read.csv("test.csv")
test$index <- 1:nrow(test)

num_groups = 8

totest <- test %>%
  group_by((row_number()-1) %/% (n()/num_groups)) %>%
  nest %>% pull(data)

head(test)
```

```
##             datetime season holiday workingday weather  temp  atemp humidity
## 1 2012-06-30 1:00:00      3       0          0       3 26.24 28.790       89
## 2 2012-06-30 2:00:00      3       0          0       2 26.24 28.790       89
## 3 2012-06-30 3:00:00      3       0          0       2 26.24 28.790       89
## 4 2012-06-30 4:00:00      3       0          0       2 25.42 27.275       94
## 5 2012-06-30 5:00:00      3       0          0       1 26.24 28.790       89
## 6 2012-06-30 6:00:00      3       0          0       1 26.24 28.790       89
##   windspeed casual registered index
## 1   15.0013      3         55     1
## 2    0.0000      7         54     2
## 3    0.0000      3         20     3
## 4    0.0000      3         15     4
## 5   11.0014      3          7     5
## 6   11.0014      6         36     6
```

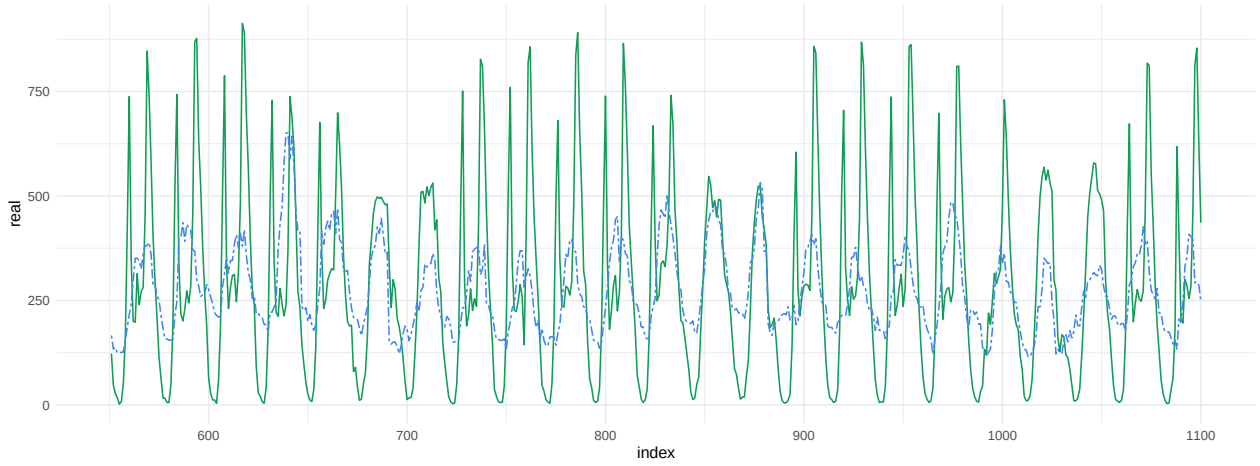
```
plots <- list()  # new empty list
for (i in 1:8) {
```

```
  p1 = dopred(data.frame(totest[i]))
  plots[[i]] <- p1  # add each plot into plot list
}
plots
```

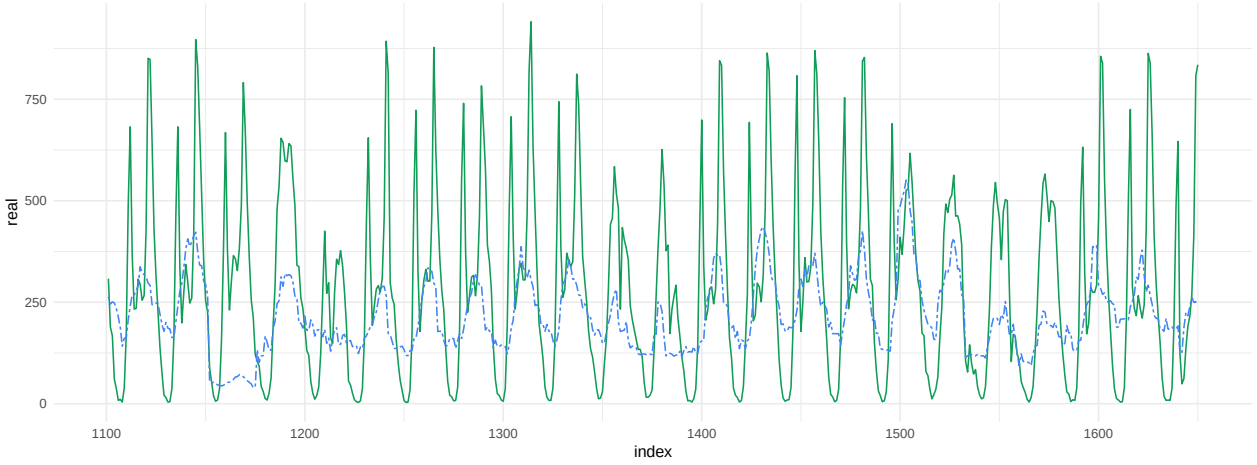
```
## [[1]]
```



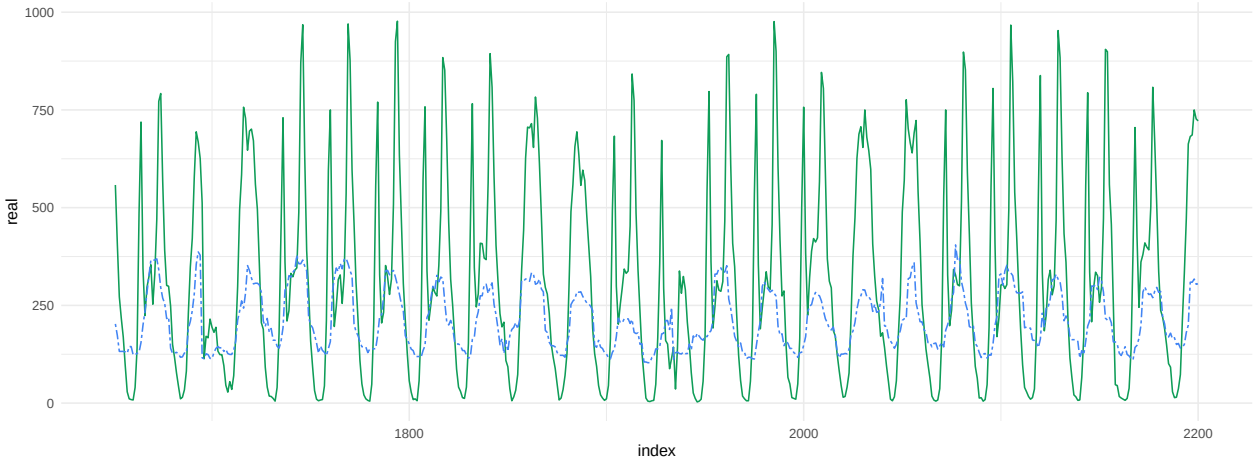

```
##
## [[2]]
```



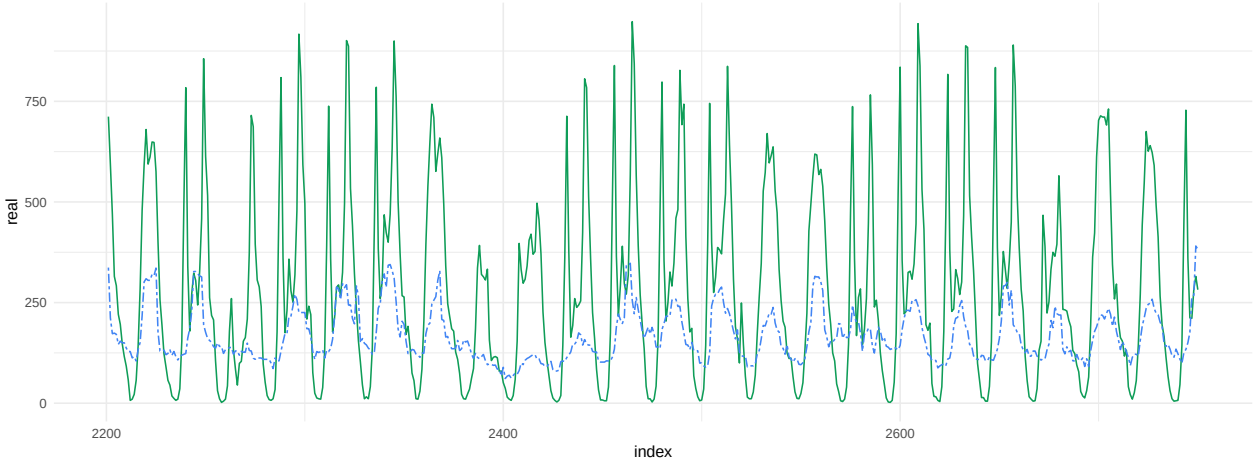

```
##
## [[3]]
```


real
750
500
250
0
1100
1200
1300
1400
1500
1600
index

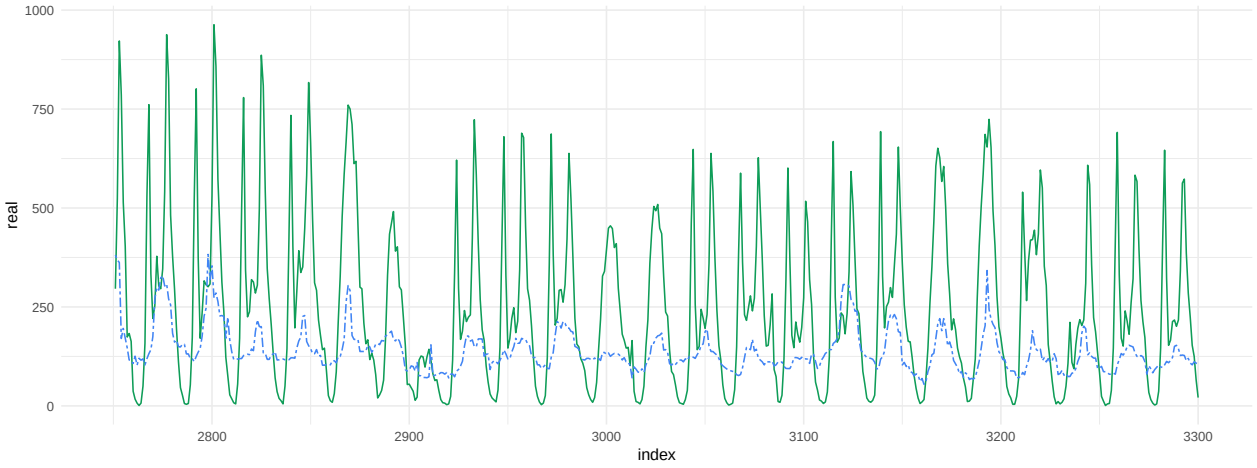
##
[[4]]


real
1000
750
500
250
0
1800
2000
2200
index

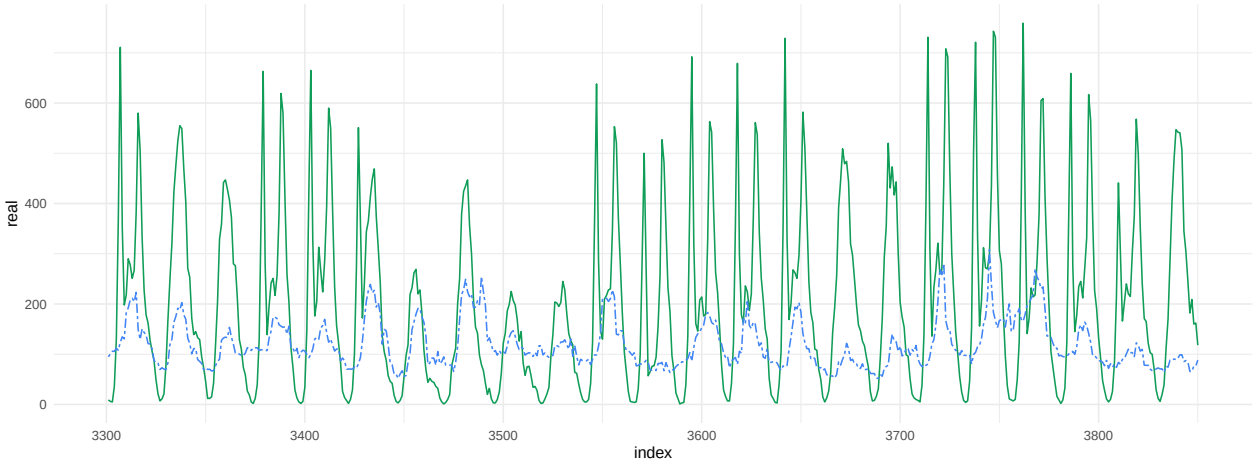
##
[[5]]


real
750
500
250
0
2200
2400
2600
index

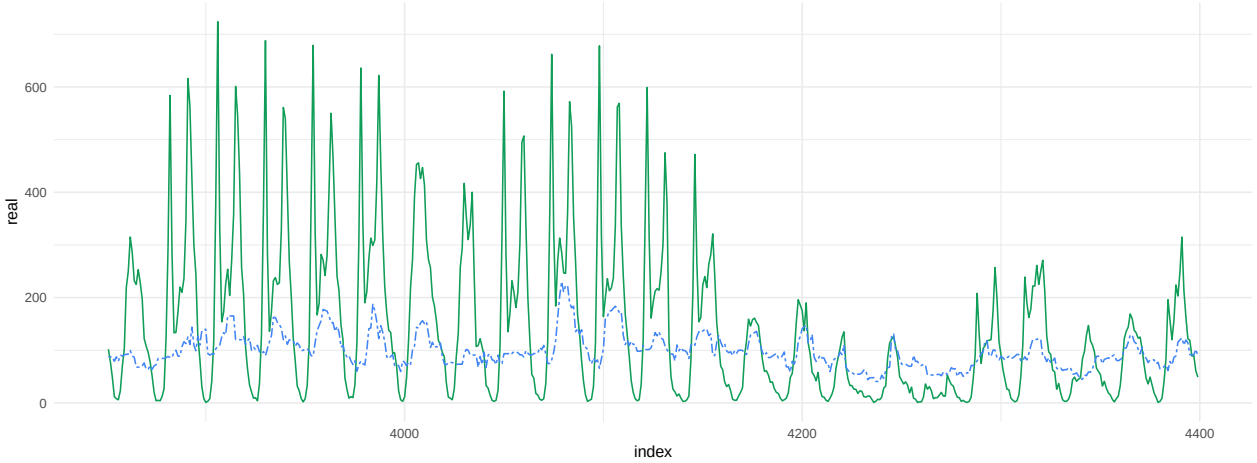
##
[[6]]


1000
750
500
250
0
real
2800
2900
3000
3100
3200
3300
index

```
##
## [[7]]
```


600
400
200
0
real
3300
3400
3500
3600
3700
3800
index

```
##
## [[8]]
```


600
400
200
0
real
4000
4200
4400
index